ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

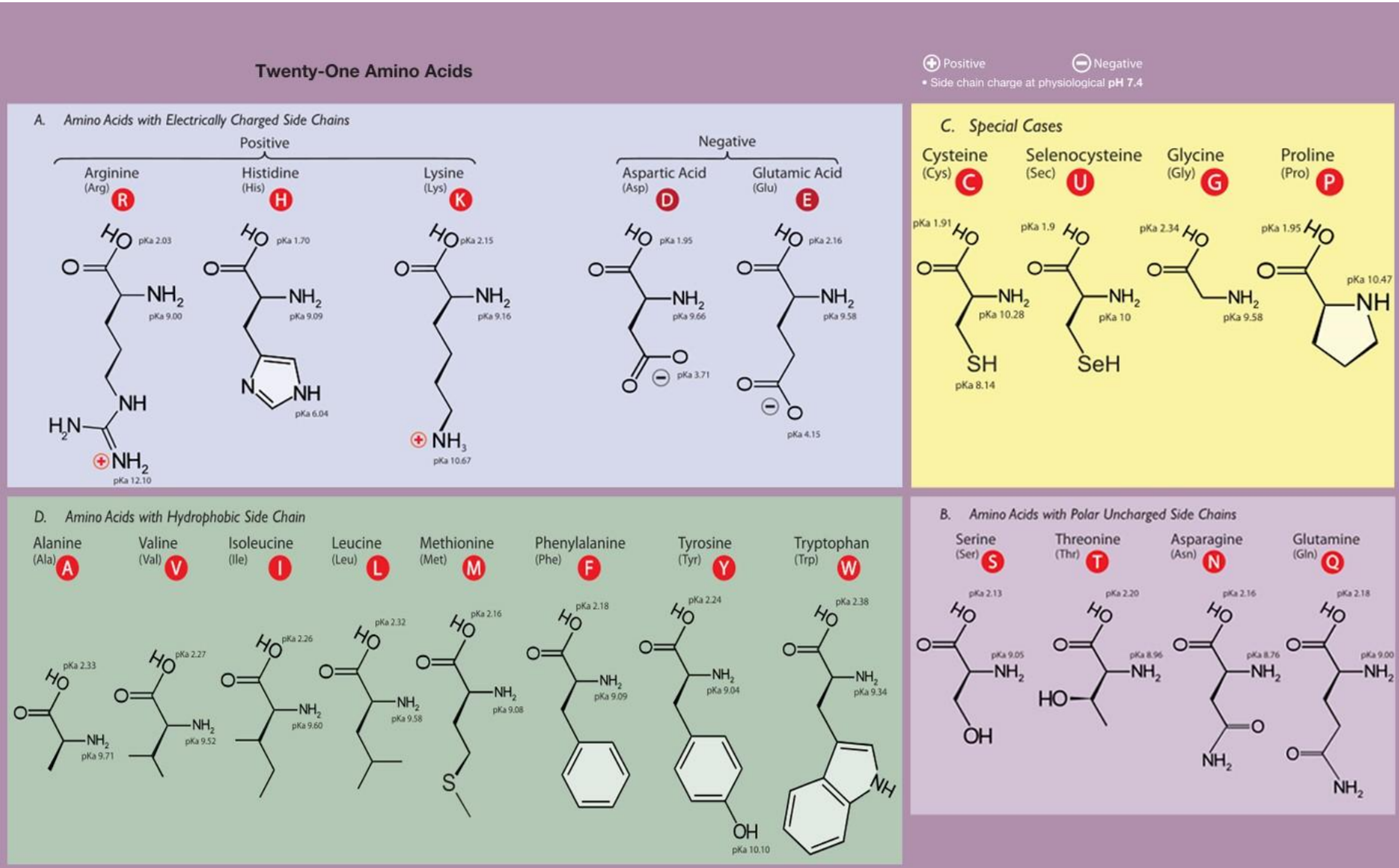
Лекция №15

Структурная биоинформатика

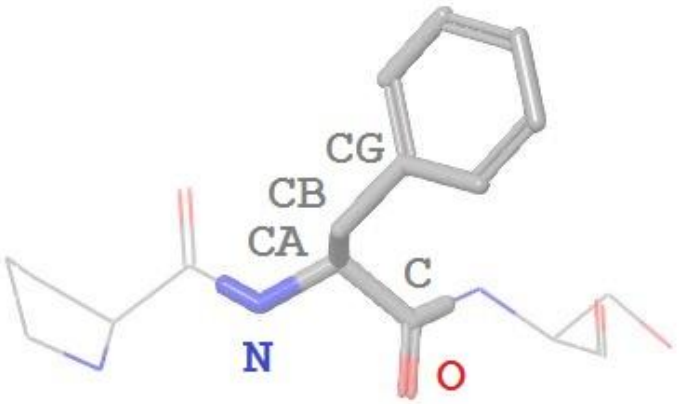
Новоселецкий Валерий Николаевич
к.ф.-м.н., доц. каф. биоинженерии
valery.novoseletsky@yandex.ru

Сайт курса http://intbio.org/bioinf2018-2019

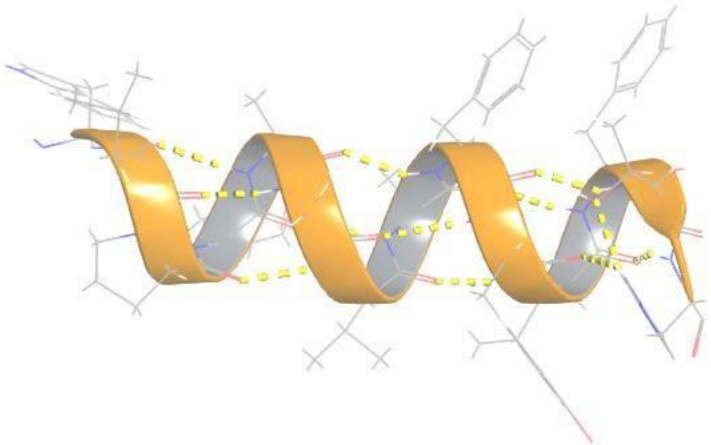
Структура аминокислот



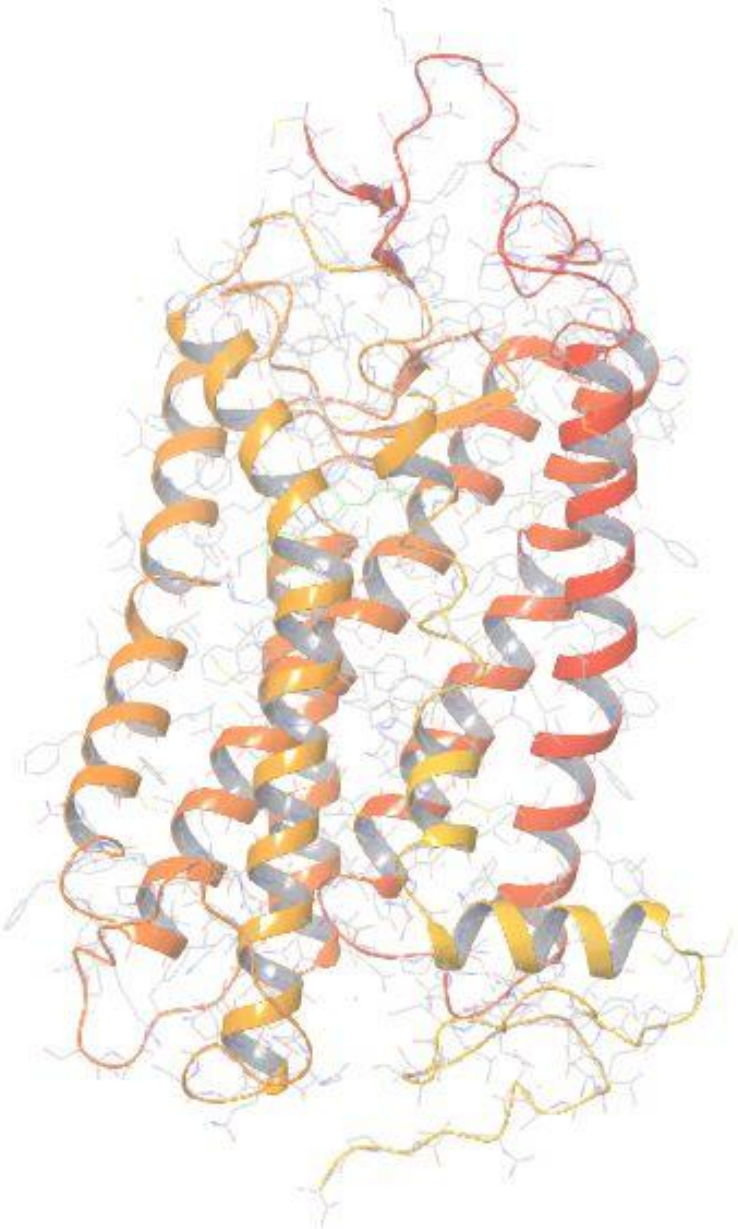
Структурная организация белков



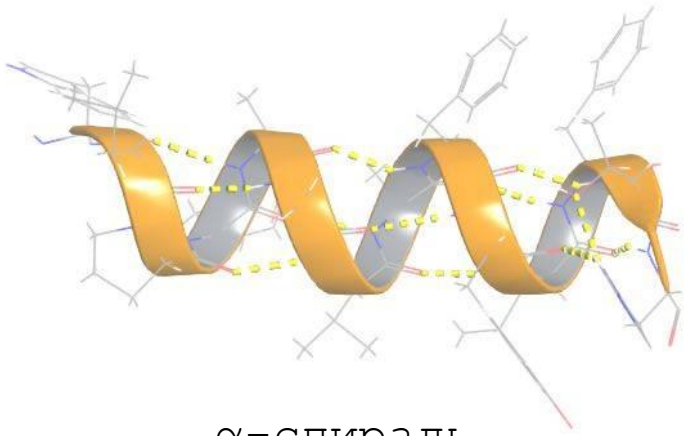
...FYV**PFS**NKT...



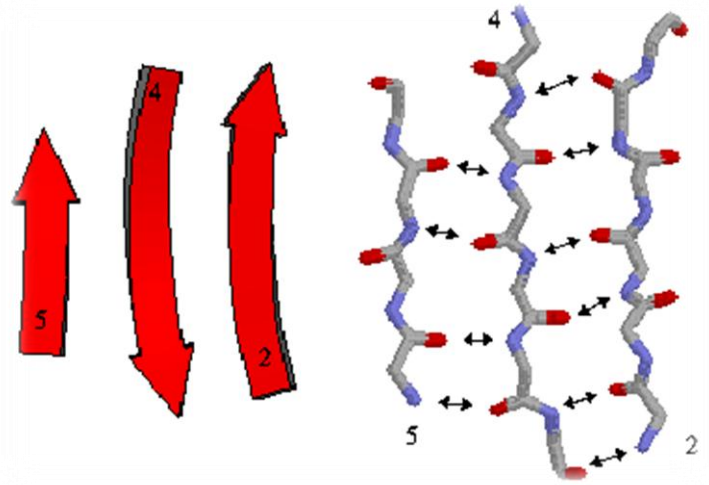
...**W**LPY**A**GVA**F**YI**F**TH...



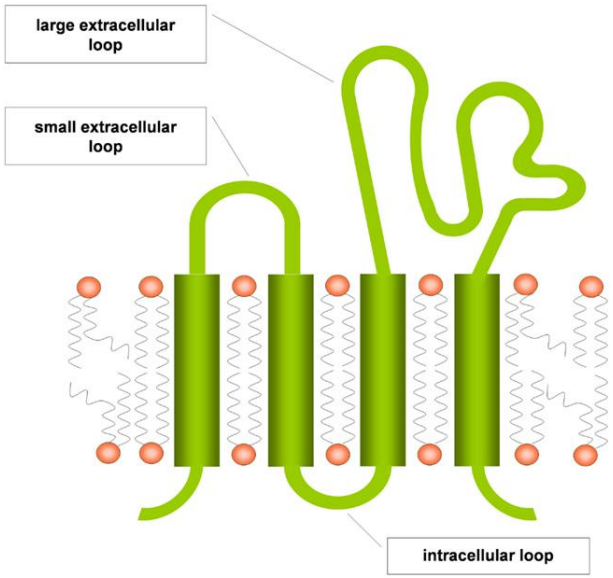
Типы вторичной структуры белков



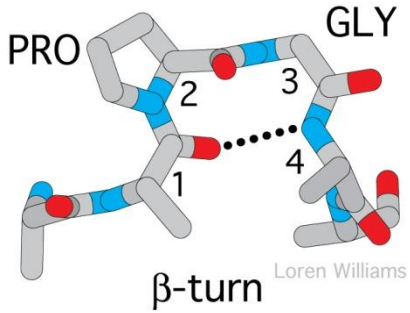
α-спираль



β-лист, состоящий из β-тяжей

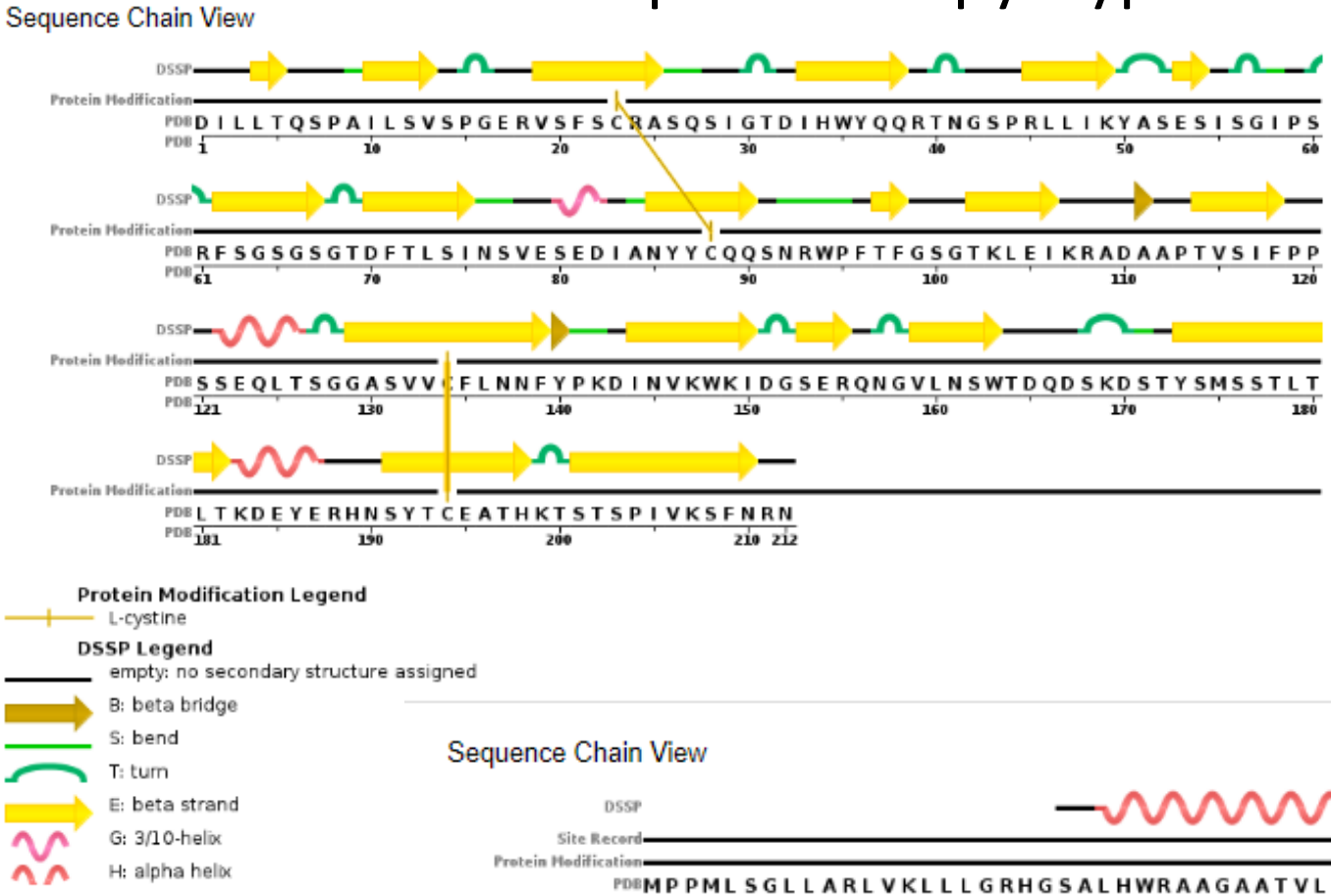


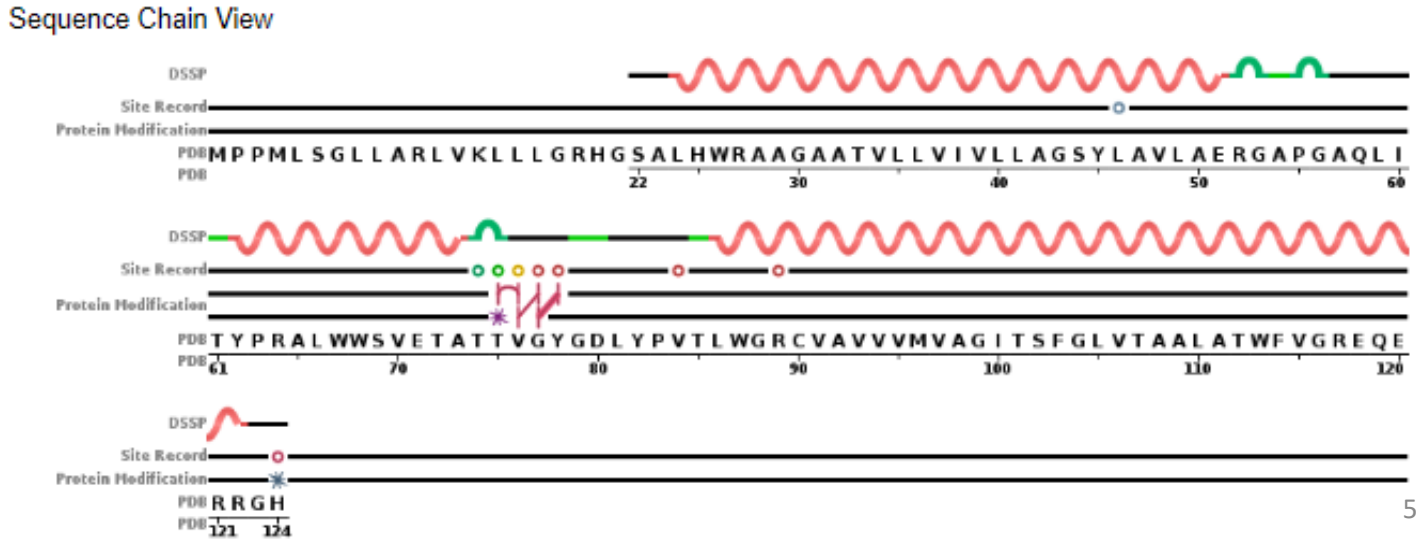
петля



поворот

Типы вторичной структуры белков



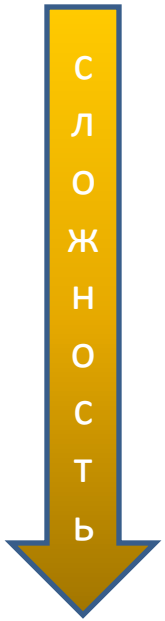


Предсказание структуры белков

Сворачивание белка в уникальную конформацию наводит на мысль об алгоритме формирования структуры белка по его последовательности, но доказательством полноты и правильности нашего понимания могла бы стать его реализация в виде компьютерной программы...

Методы предсказания структуры по последовательности:

- Предсказание вторичной структуры;
- Моделирование по гомологии;
- Распознавание типов укладки (по известной библиотеке фолдов);
- Априорное предсказание новых типов укладки.



Предсказание вторичной структуры

В настоящее время предсказание может быть выполнено с помощью различных алгоритмов и является относительно точным - ~ 80% элементов вторичной структуры выявляется правильно.

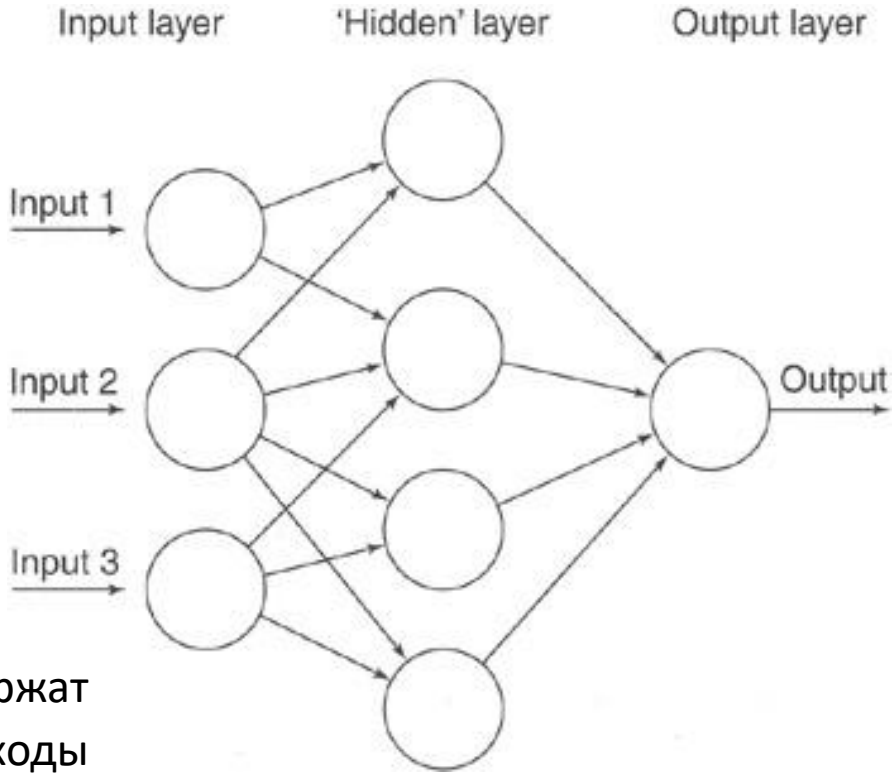
```
                    10        20        30        40        50
AA sequence     ALVEDPPLKVSEGGLIREGYDPDLDALRAAHREGVAYFLELEERERERTG
Prediction      HH----------EEE------HHHHHHHHHH-HHHHHHHHHHHHHHH
Experiment      -E----------E-----HHHHHHHHHHHHHHHHHHHHHHHHHHHH-

                    60        70        80        90        100
AA sequence     IPTLKVGYMAVFGYYLEVTRPYYERVPKEYRPVQTLKDRQRYTLPEMKEK
Prediction      --EEEEEEEEEEEEEEE----------EEEEEEEE--EEEE-HHHHHH
Experiment      ----EEEEE---EEEEEEEHHHHHH-----EEEEE---EEEEE-HHHHHH

                    110       120
AA sequence     EREVYRLEALIRRREEEVFLEVRERAKRQ
Prediction      HHHHHHHHHHHHHHHHHHHHHHHHHHHH
Experiment      HHHHHHHHHHHHHHHHHHHHHHHHHHH--
```

Наиболее мощные методы предсказания вторичной структуры основаны на нейронных сетях.

Нейронные сети

Искусственные нейронные сети (ИНС) — математические модели, а также их программные или аппаратные реализации, построенные по принципу организации и функционирования сетей нервных клеток живого организма.

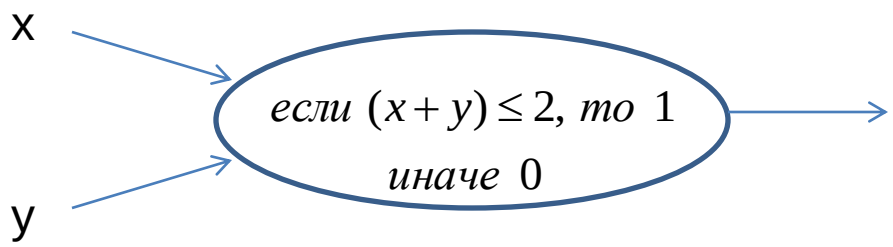
В вычислительной схеме одиночный нейрон является вершиной графа с несколькими входящими ребрами и одним исходящим. Для формирования сети необходимо соединить выходы одних нейронов со входами других.

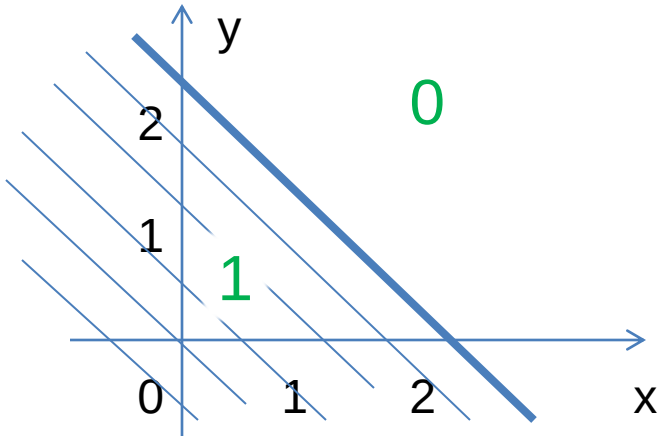


При этом некоторые нейроны содержат входы для всей сети, некоторые – выходы наружу, а некоторые с внешним миром не связаны (скрытые нейроны).

Нейронные сети. Геометрическая интерпретация

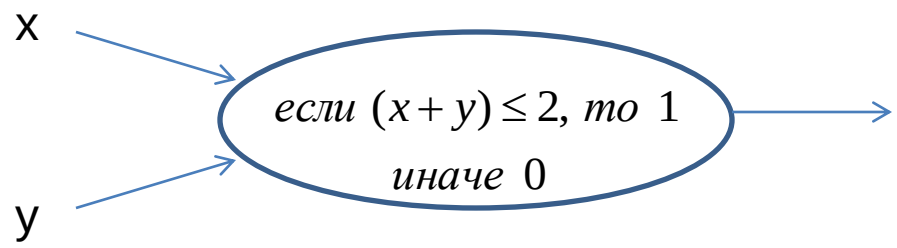
Если интерпретировать пару чисел (x, y) на входе как точку на плоскости, то данный нейрон принимает решение, на какой стороне от линии находится вход.

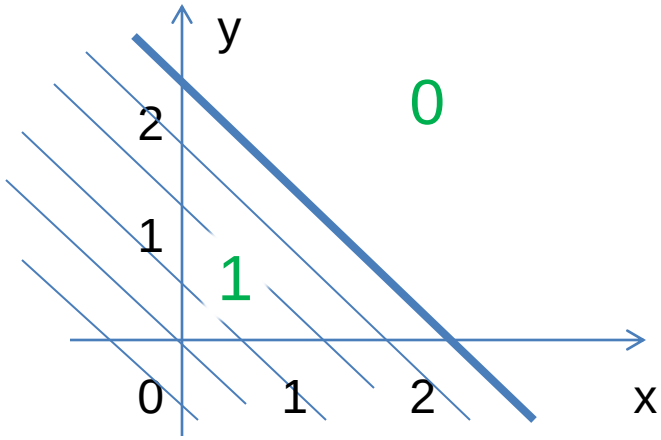


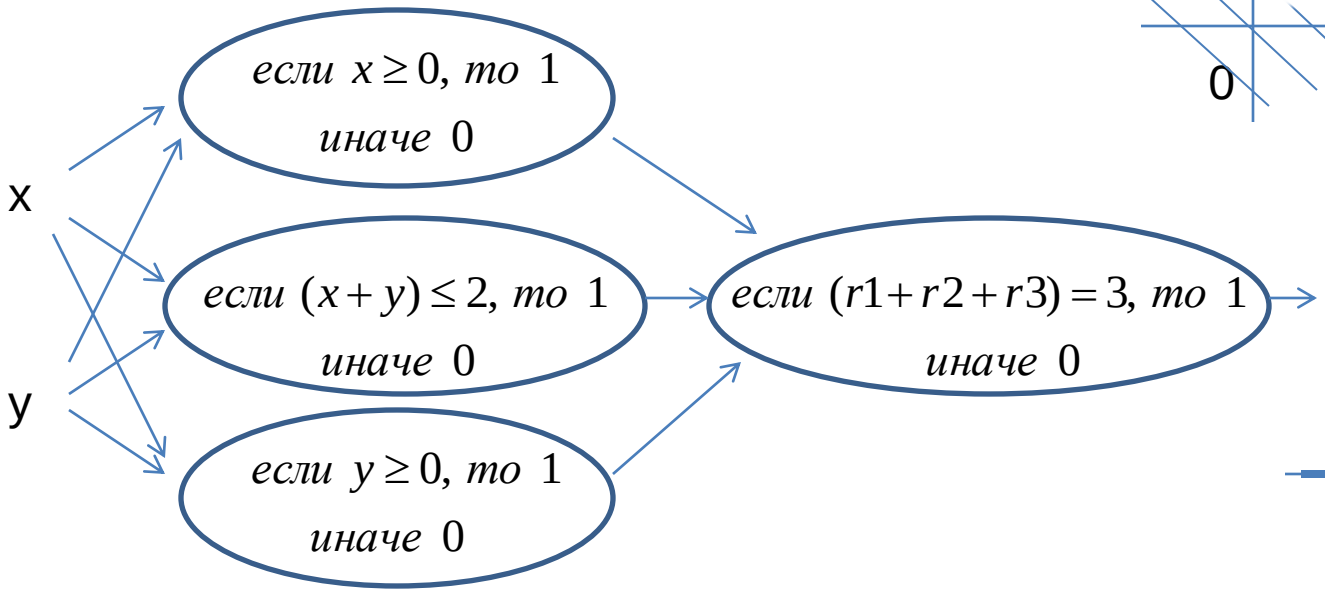


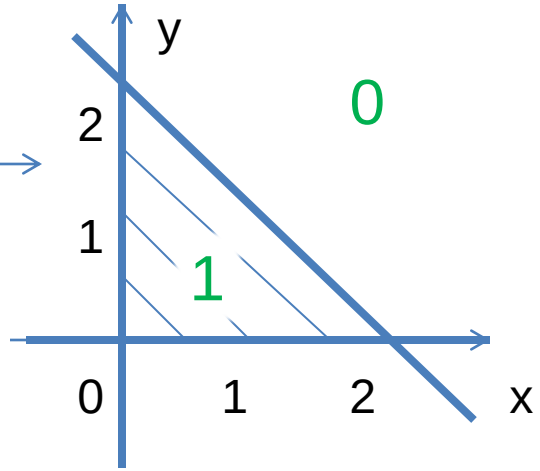
Нейронные сети. Геометрическая интерпретация

Если интерпретировать пару чисел (x, y) на входе как точку на плоскости, то данный нейрон принимает решение, на какой стороне от линии находится вход.



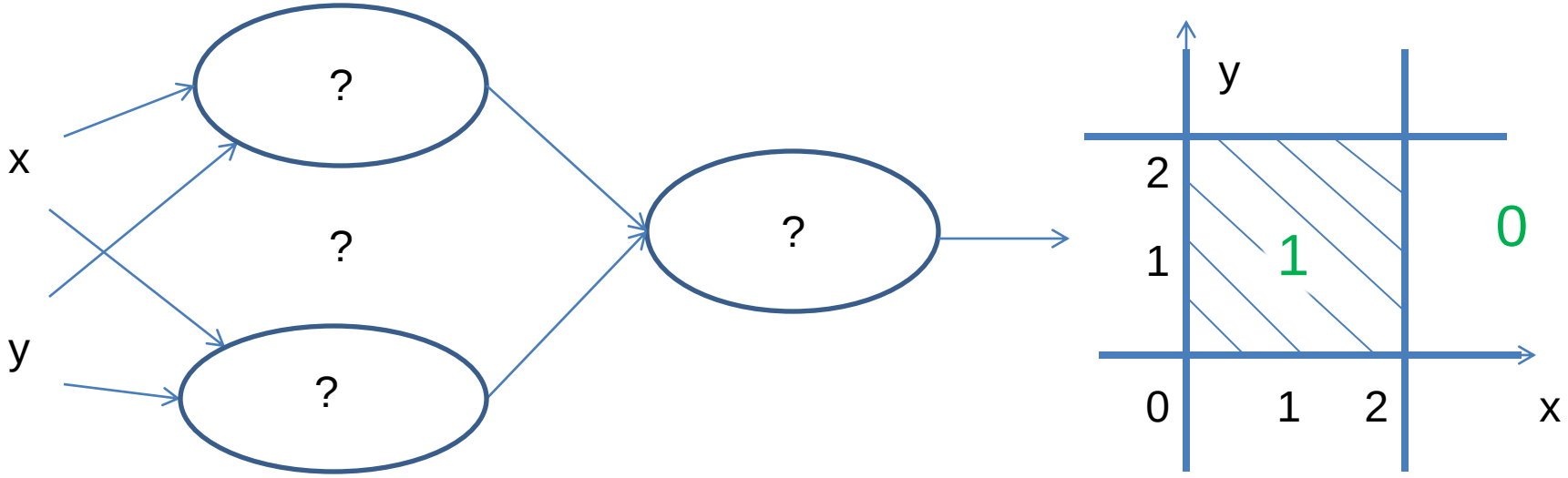




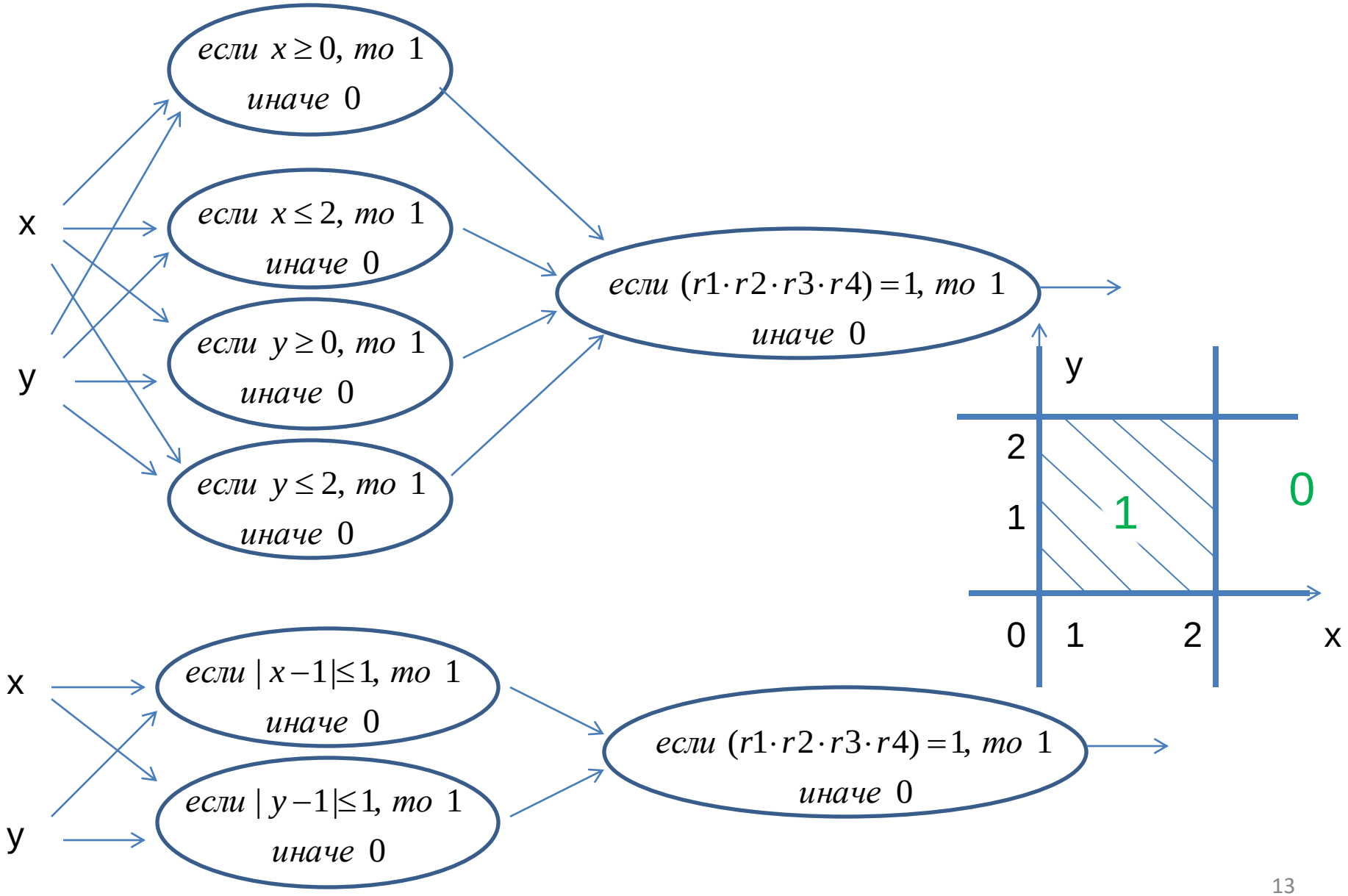


Нейронная сеть определяется топологией связей, весами и формулой принятия решения в узлах. Очевидно, сеть может принимать более сложные решения, чем один нейрон.

Нейронные сети. Геометрическая интерпретация

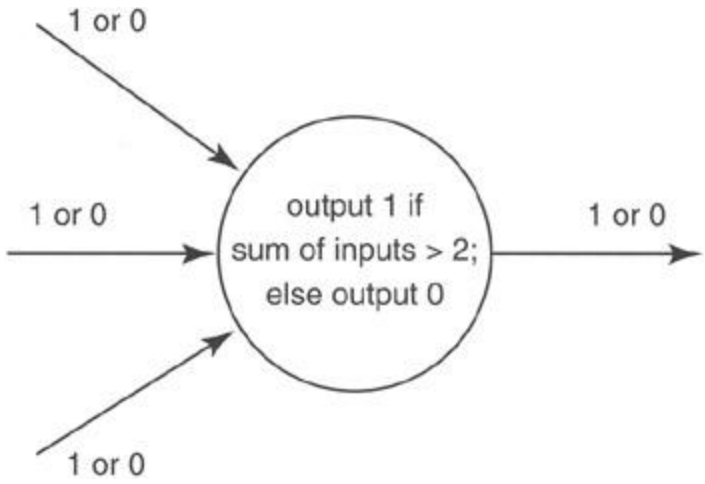


Нейронные сети. Геометрическая интерпретация



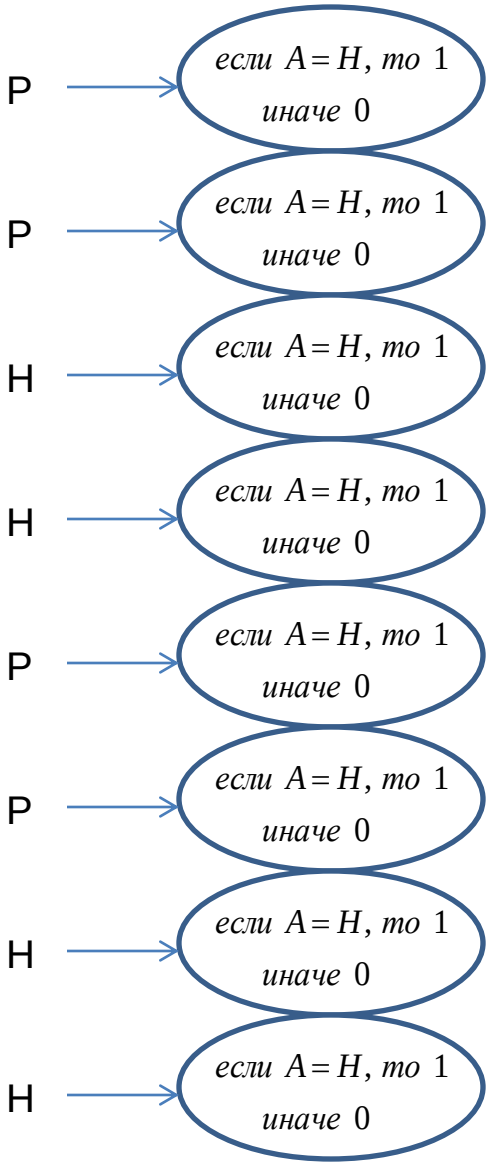
Нейронные сети. Веса связей

Неограниченная сложность возможна как при создании и соединении нейронов, так и при определении строгости связей. Если вместо того, чтобы просто просуммировать входные сигналы ***i1 + i2 + i3***, использовать взвешенную сумму входов **10*i1 + i2 + 0,5*i3**, то сеть станет более чувствительной ко входу 1 и менее ко входу 3.



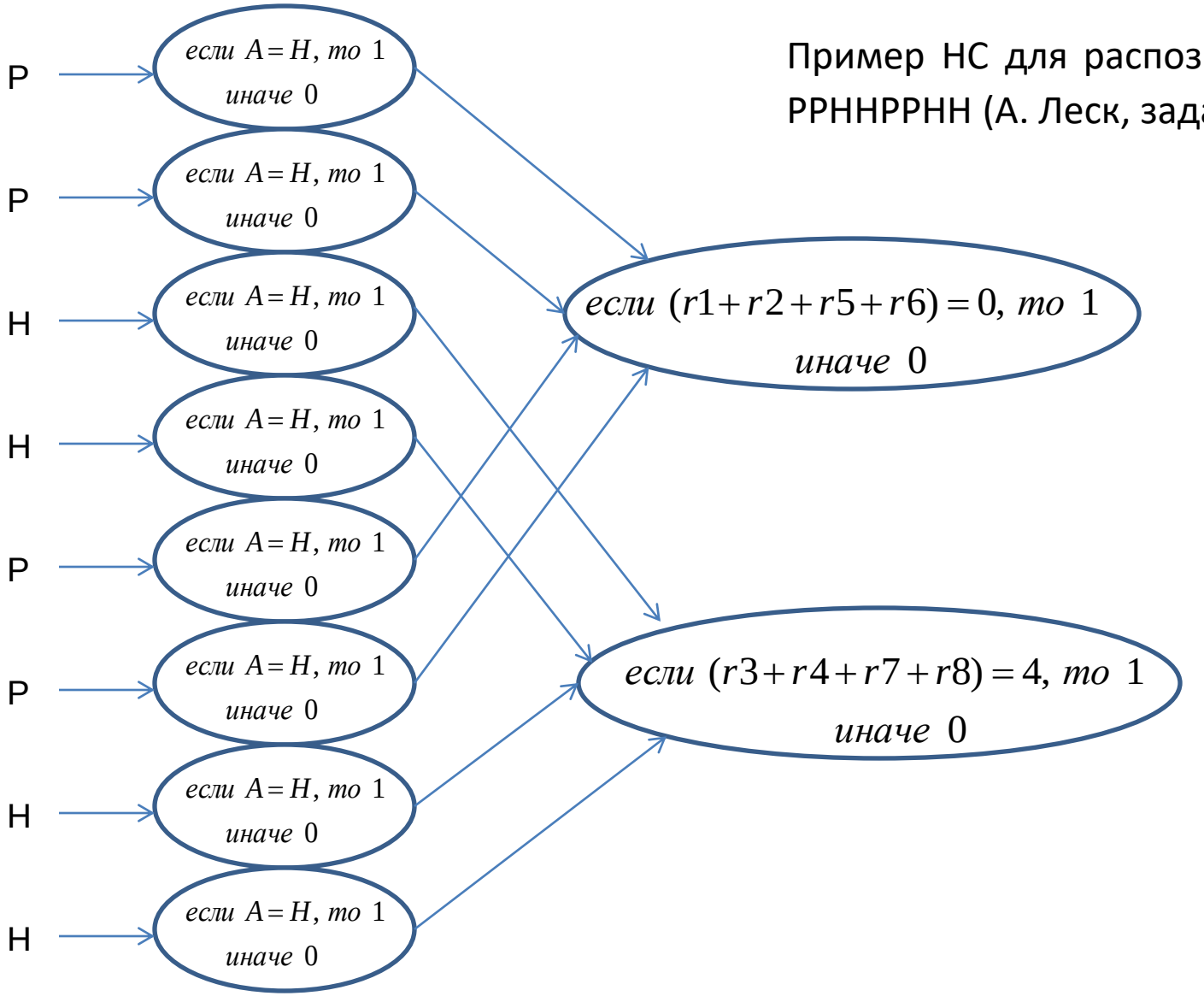
В процессе обучения происходит подбор параметров при неизменной топологии сети. Для этого применяют сеть с начальными параметрами к различным примерам и сравнивают ответ с правильным. При несовпадении производят уточнение параметров.

Нейронные сети. Распознавание α-спирали



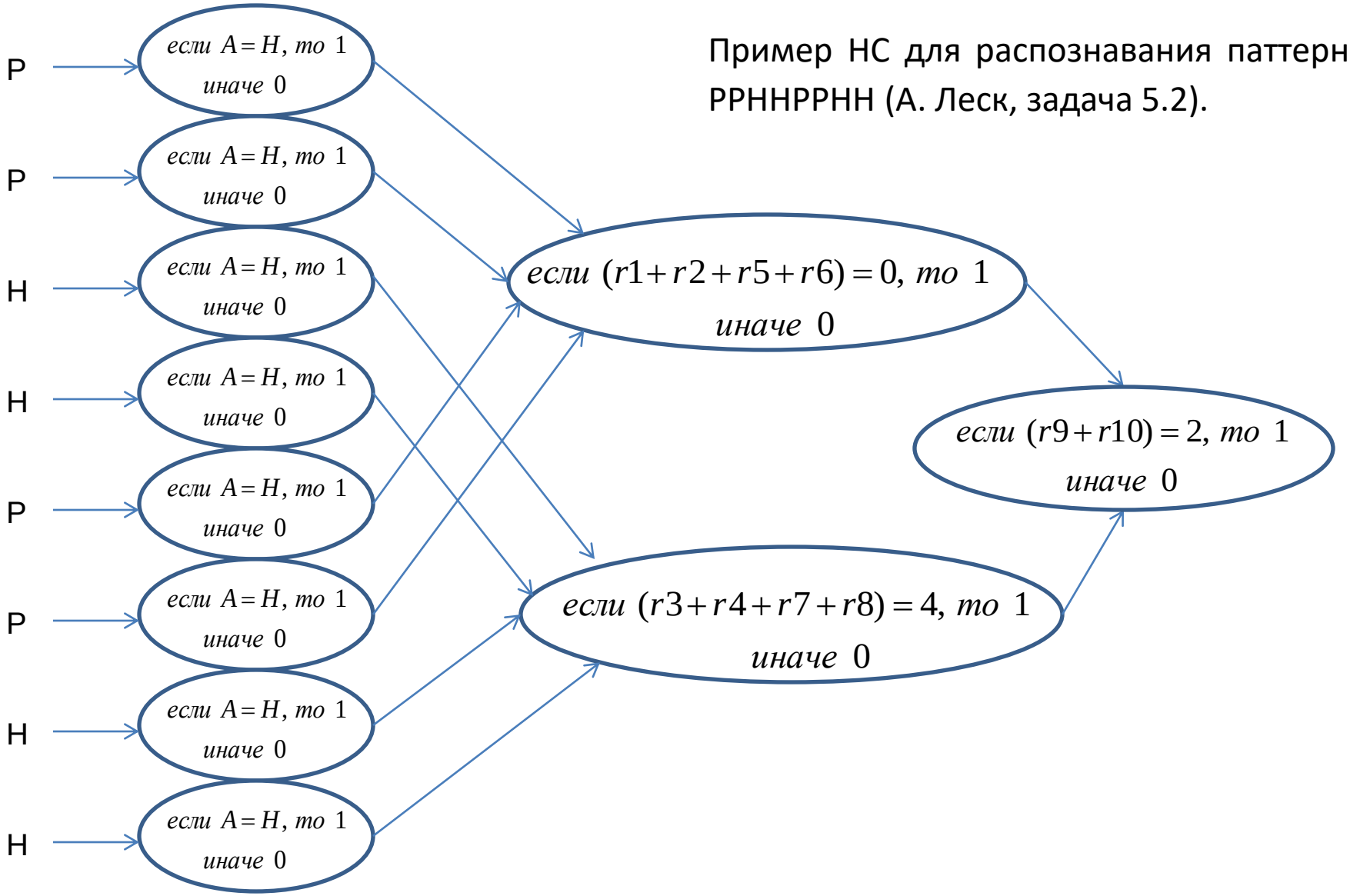
Пример НС для распознавания паттерна PPHHPPHH (А. Леск, задача 5.2).

Нейронные сети. Распознавание α-спирали



Пример НС для распознавания паттерна PPHHPPHH (А. Леск, задача 5.2).

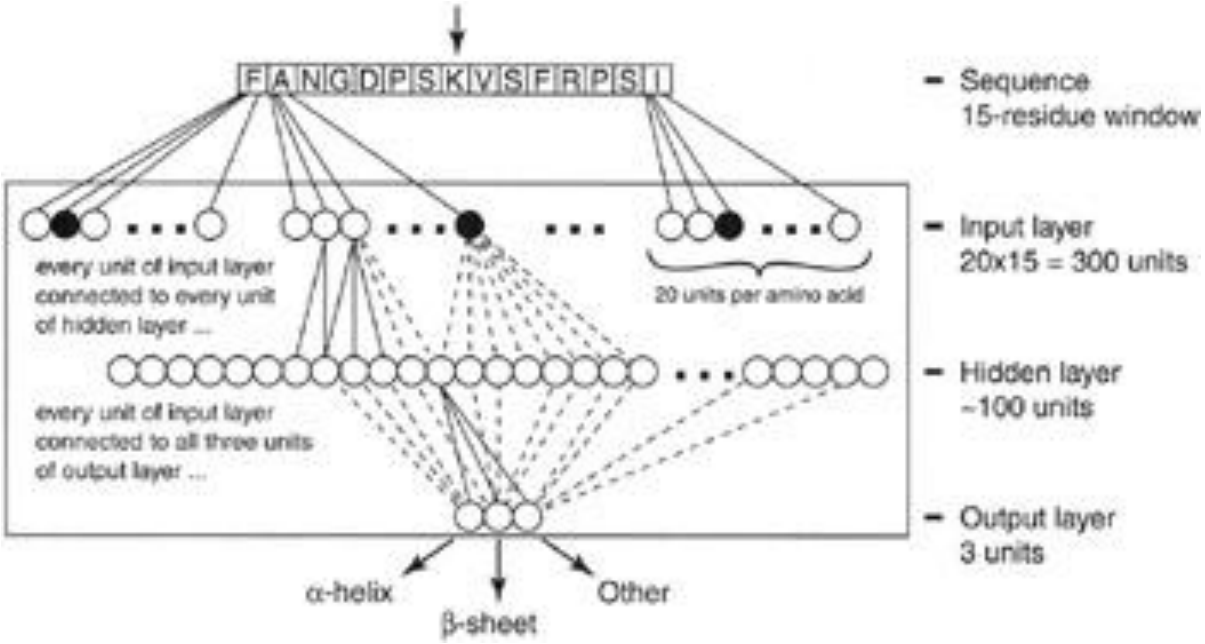
Нейронные сети. Распознавание α-спирали



Пример НС для распознавания паттерна PPHHPPHH (А. Леск, задача 5.2).

Предсказание вторичной структуры. Нейронные сети

Пример нейронной сети для предсказания вторичной структуры (PSIPRED, 1999).



Входная область сканирует последовательность окном шириной в 15 остатков, при этом предсказание делается для центрального остатка. Каждому из остатков соответствует 20 входных нейронов, один из которых активен.

Скрытая область состоит из ~ 100 нейронов, соединенных с каждым нейроном ввода и вывода.

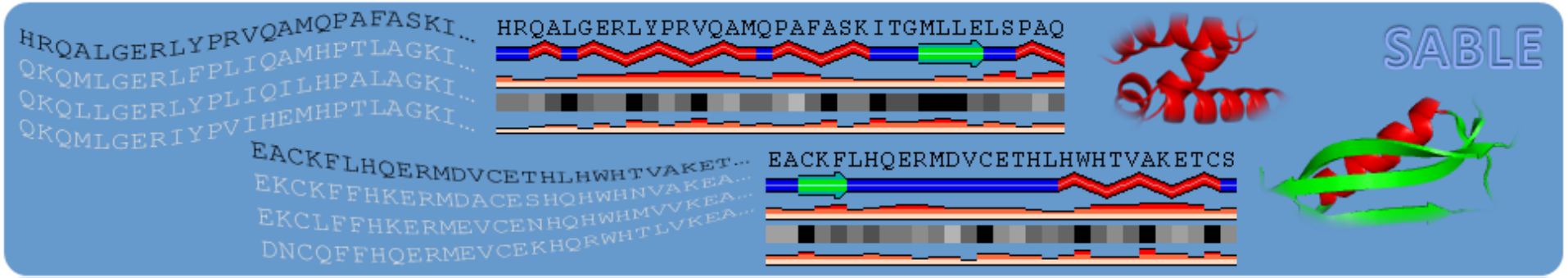
Область вывода состоит из трех нейронов, которые делают предсказание: спираль, лист или ни то, ни другое.

Предсказание вторичной структуры. PSIPRED и SABLE

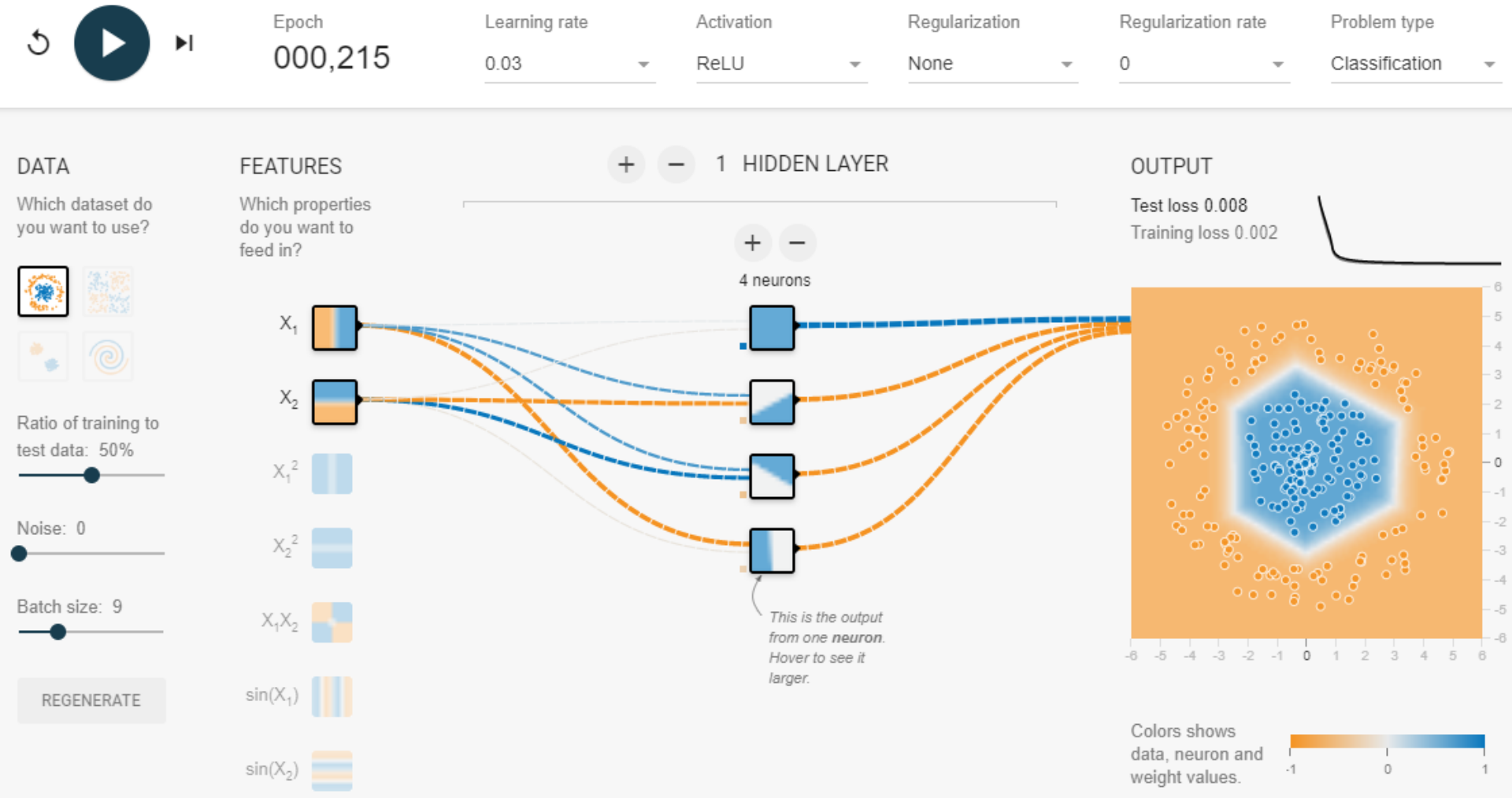
The PSIPRED Protein Sequence Analysis Workbench

The PSIPRED Protein Sequence Analysis Workbench aggregates several UCL structure prediction methods into one location. Users can submit a protein sequence, perform the predictions of their choice and receive the results of the prediction via e-mail or the web.
For a summary of the available methods you can read More...

PSIPRED

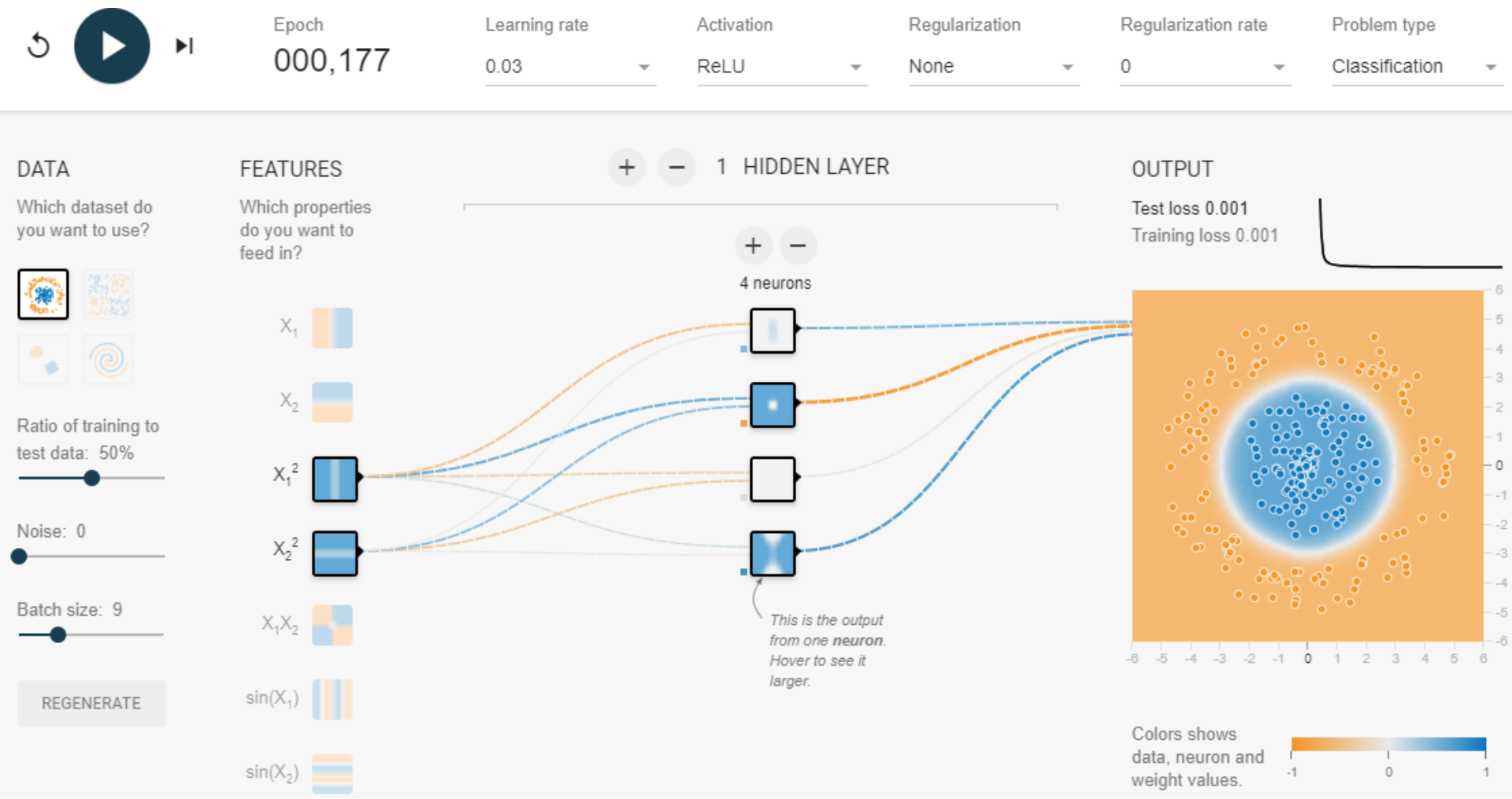


A Neural Network Playground



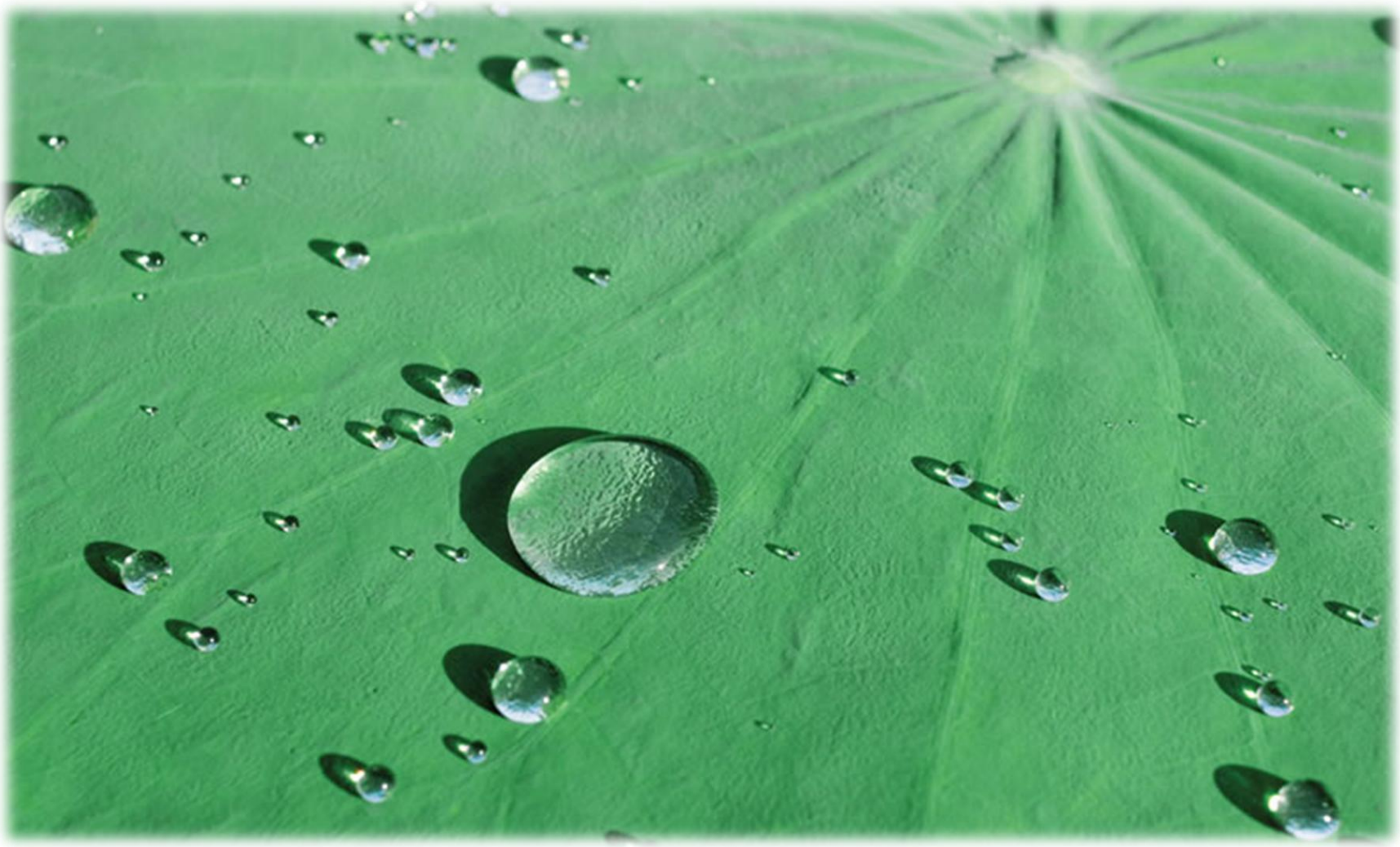
http://playground.tensorflow.org

A Neural Network Playground



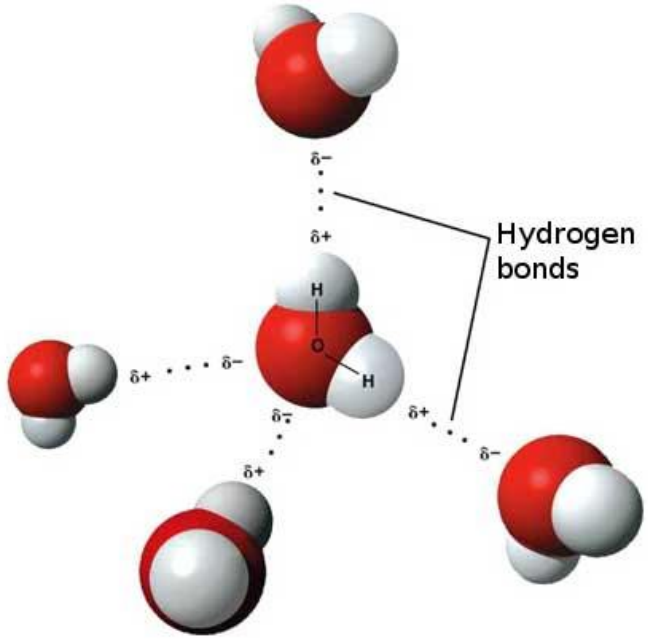
http://playground.tensorflow.org

Гидрофобность



Гидрофобность

Гидрофобный эффект – следствие большей упорядоченности молекул воды вокруг неполярной молекулы.

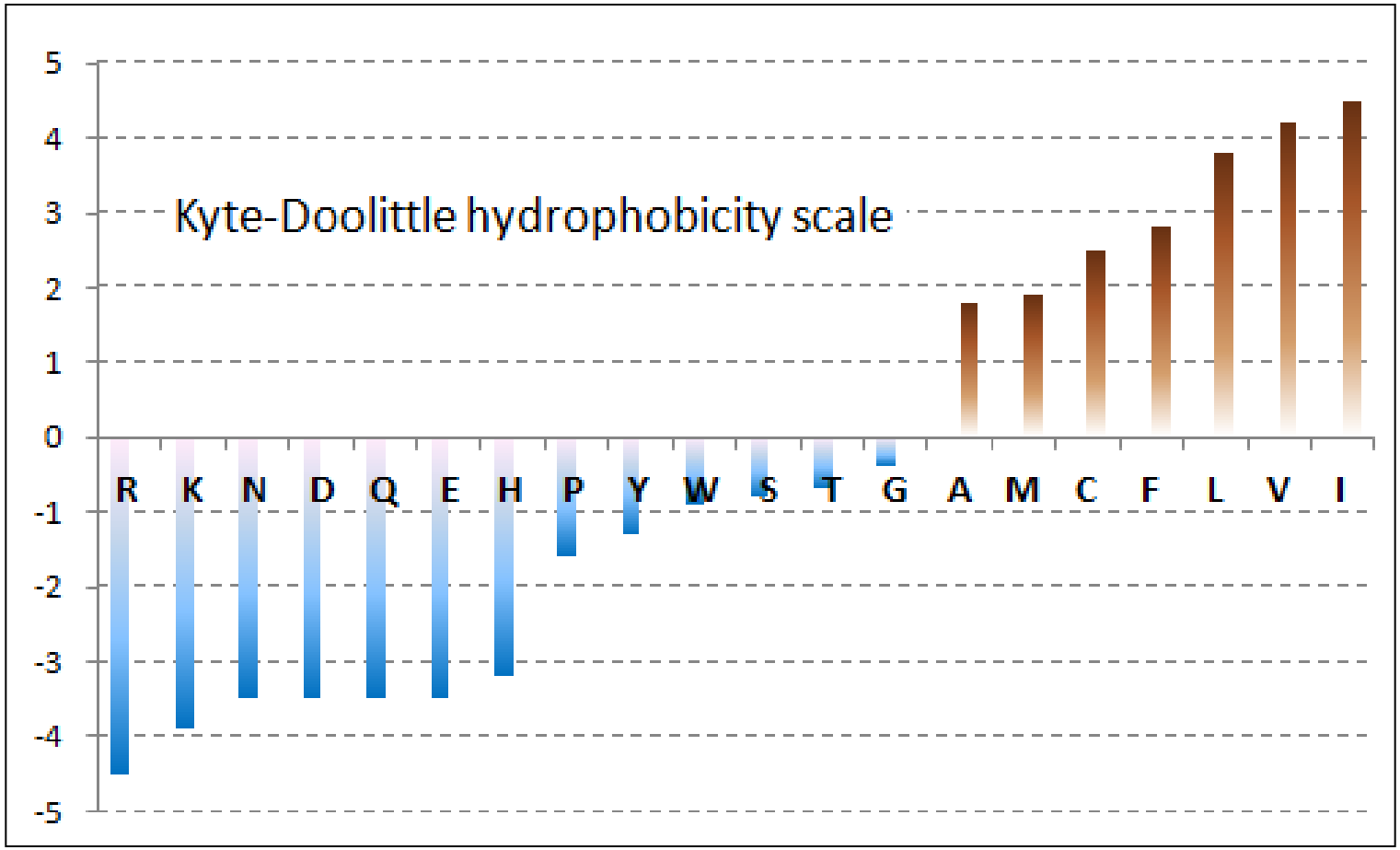


Мерой гидрофобности молекул может служить коэффициент разделения – равновесное отношение концентраций вещества в двух фазах в случае несмешивающихся растворителей:

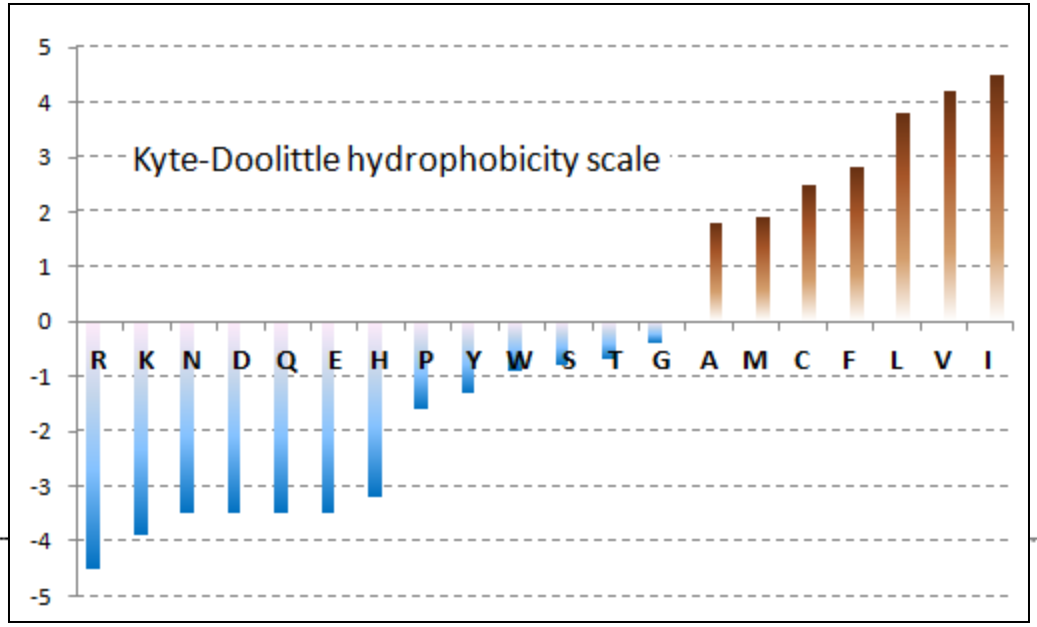
$$log\ P_{oct/wat} = log\left(\frac{[solute]_{octanol}}{[solute]^{un-ionized}_{water}}\right)$$

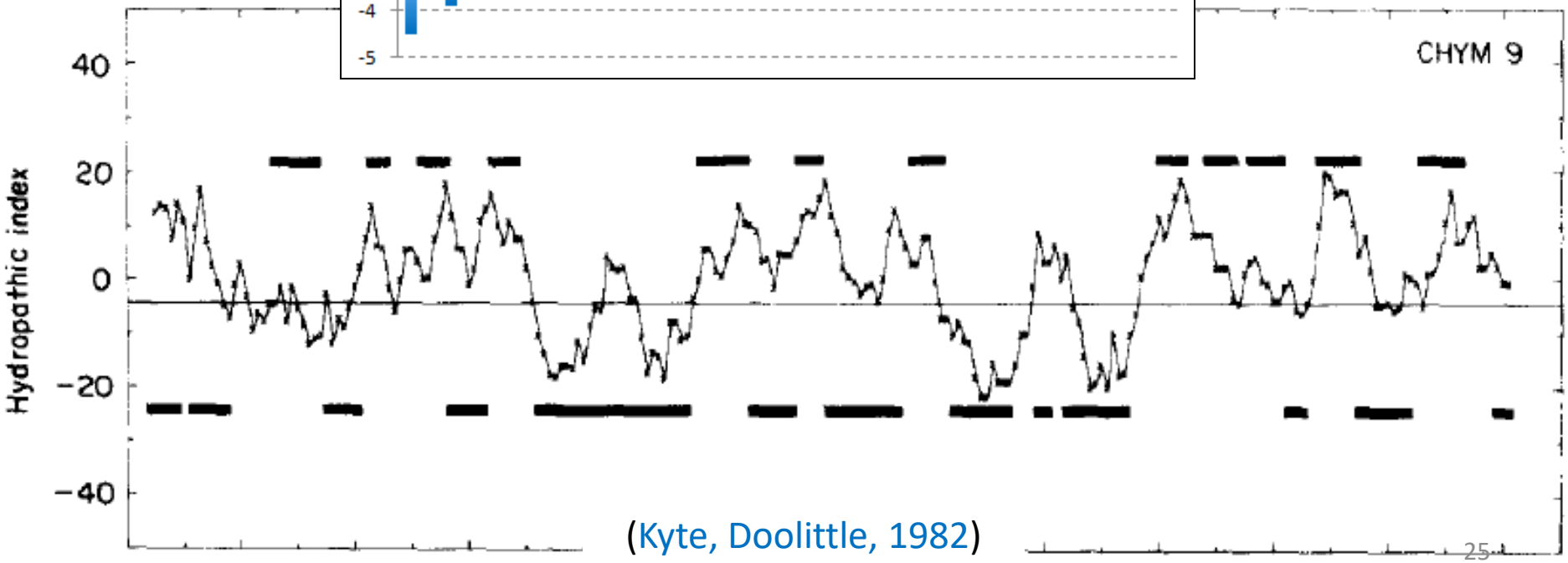
Гидрофобность

(Kyte, Doolittle, 1982) – распространенная шкала туманного происхождения



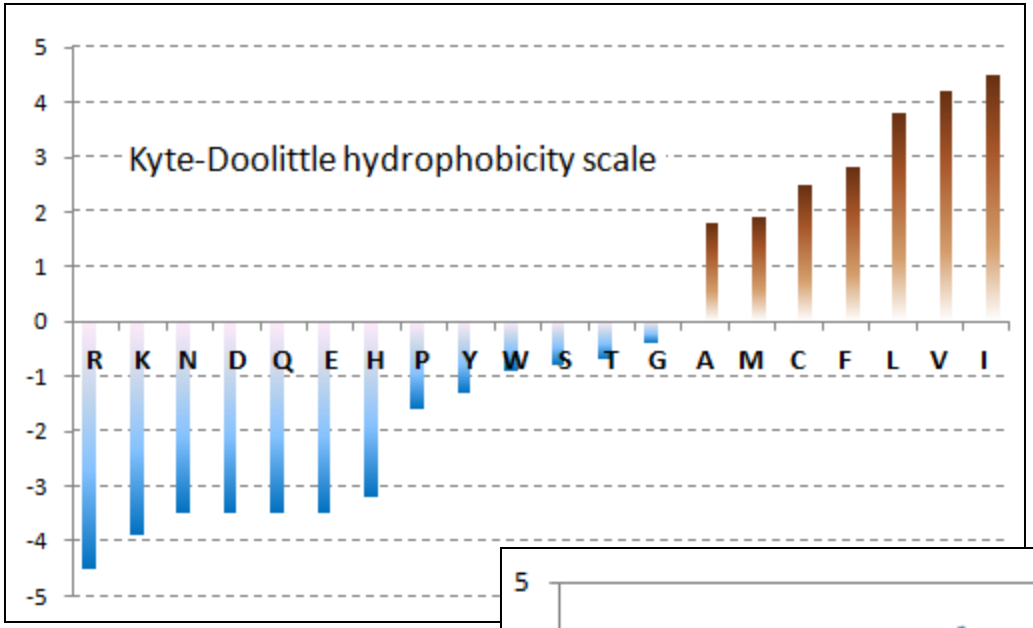
Предсказание топологии. Профили гидрофобности

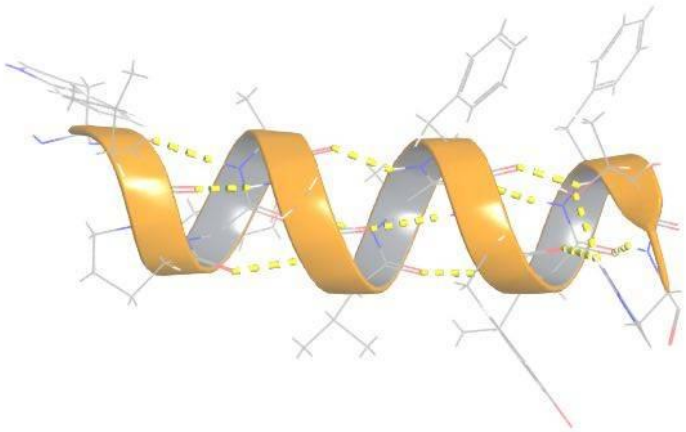




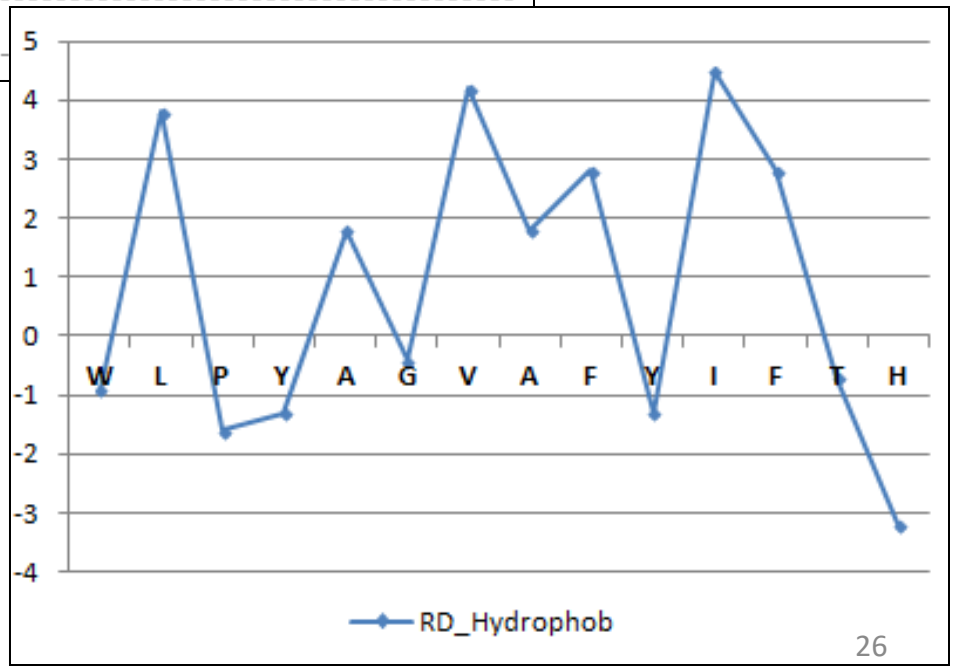
(Kyte, Doolittle, 1982)

Предсказание топологии. Профили гидрофобности

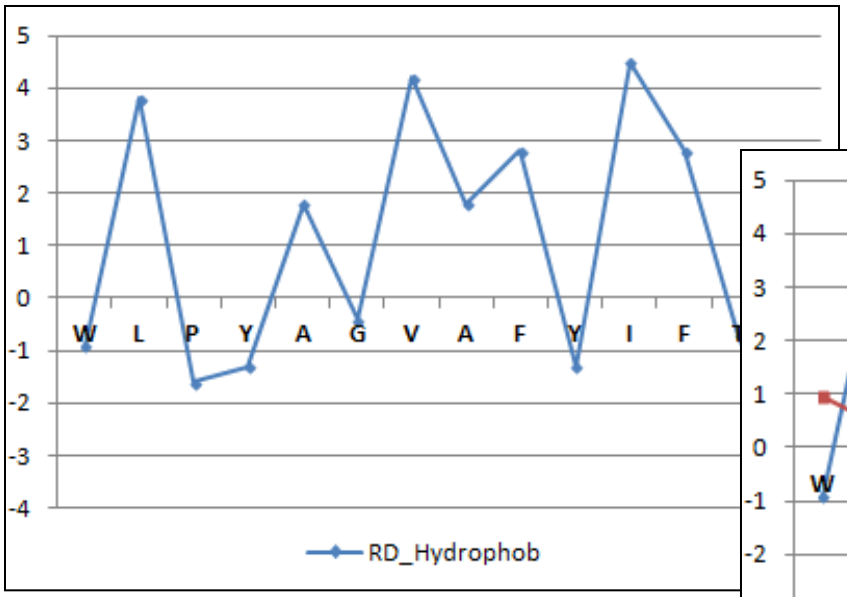


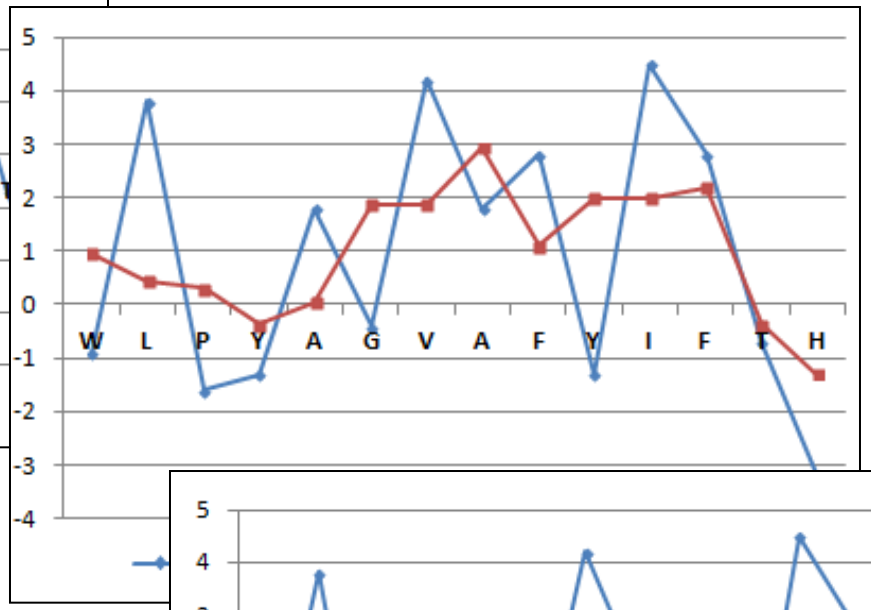


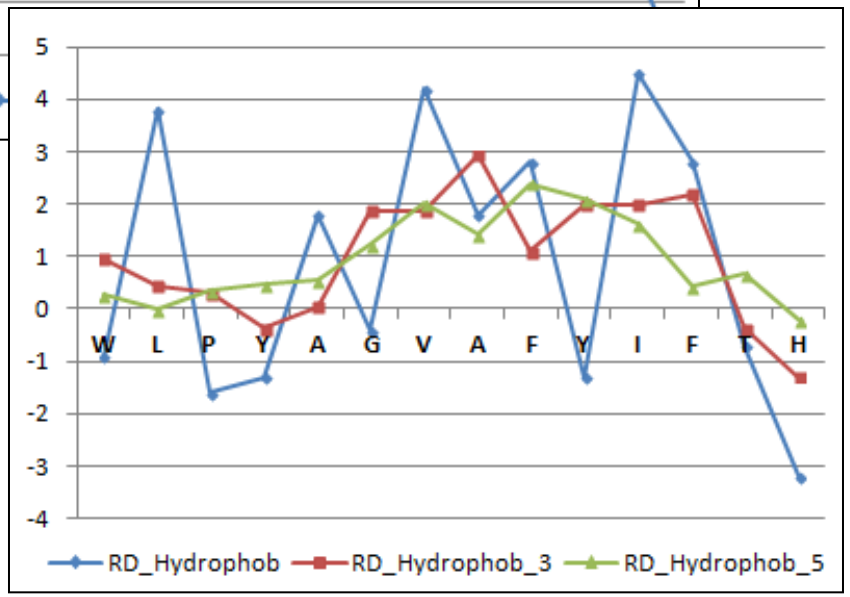
...**W**LPY**A**GVA**F**YI**F**TH...

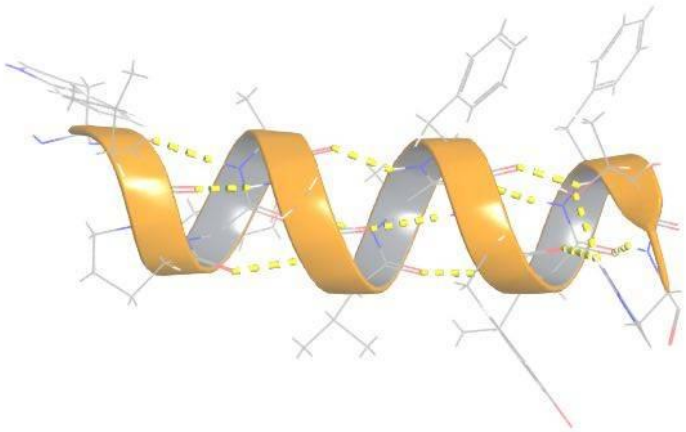


Предсказание топологии. Профили гидрофобности





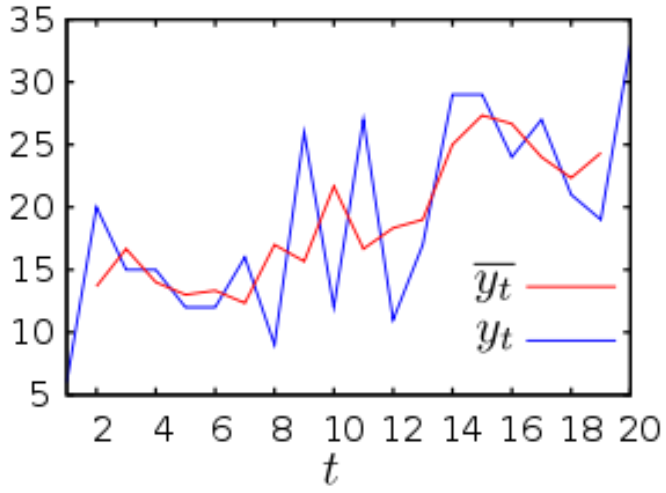




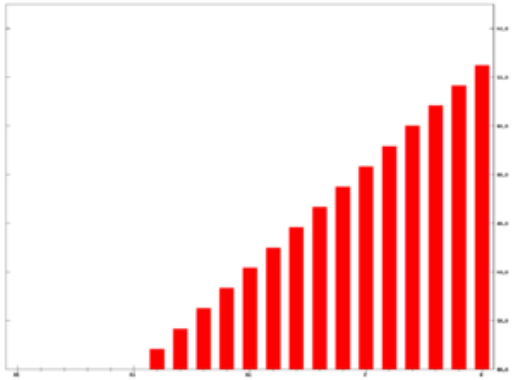
...**W**LPY**A**GVA**F**YI**F**TH...

Скользящее среднее

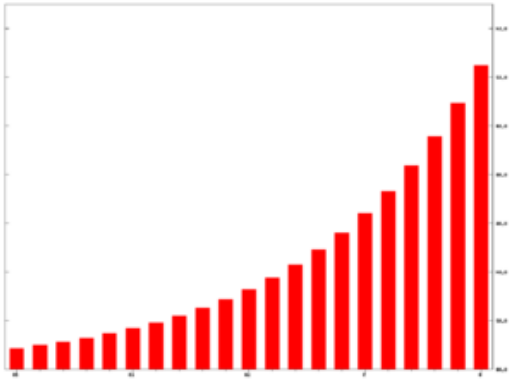
$$\bar{y}_t = \frac{y_t + y_{t-1}}{2}$$



$$WMA_t = \frac{\sum_{i=0}^{n-1} W_{t-i} \cdot p_{t-i}}{\sum_{i=0}^{n-1} W_{t-i}}$$

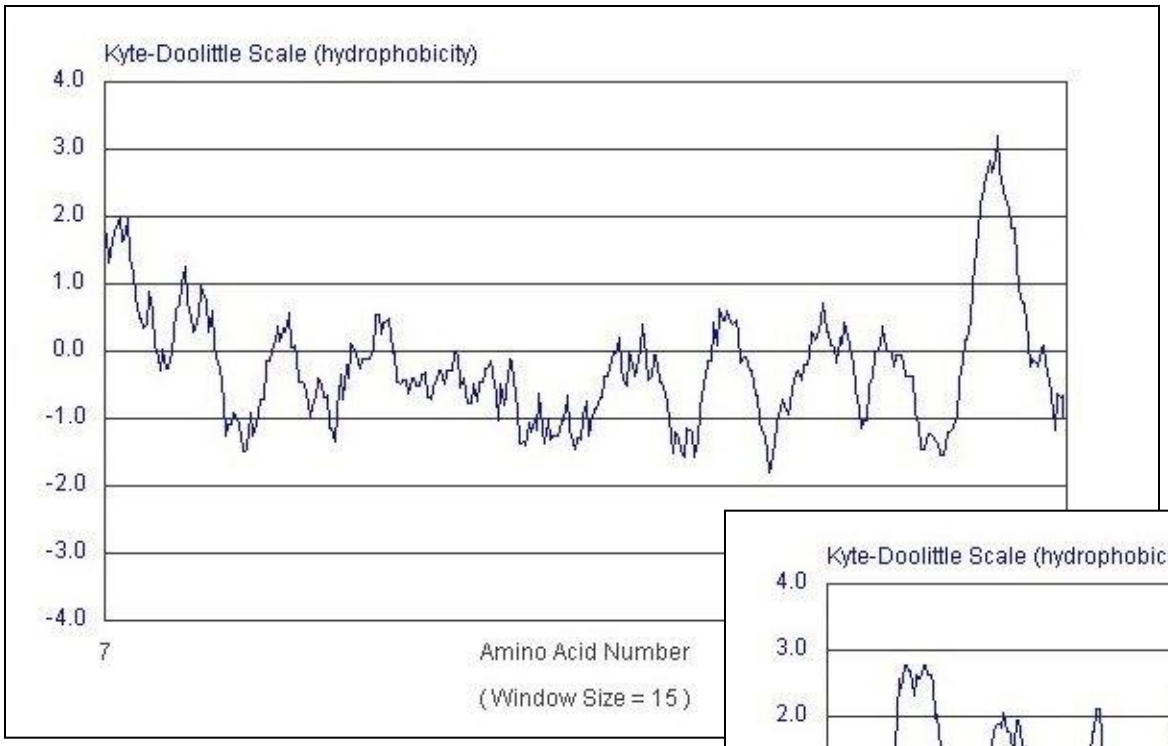


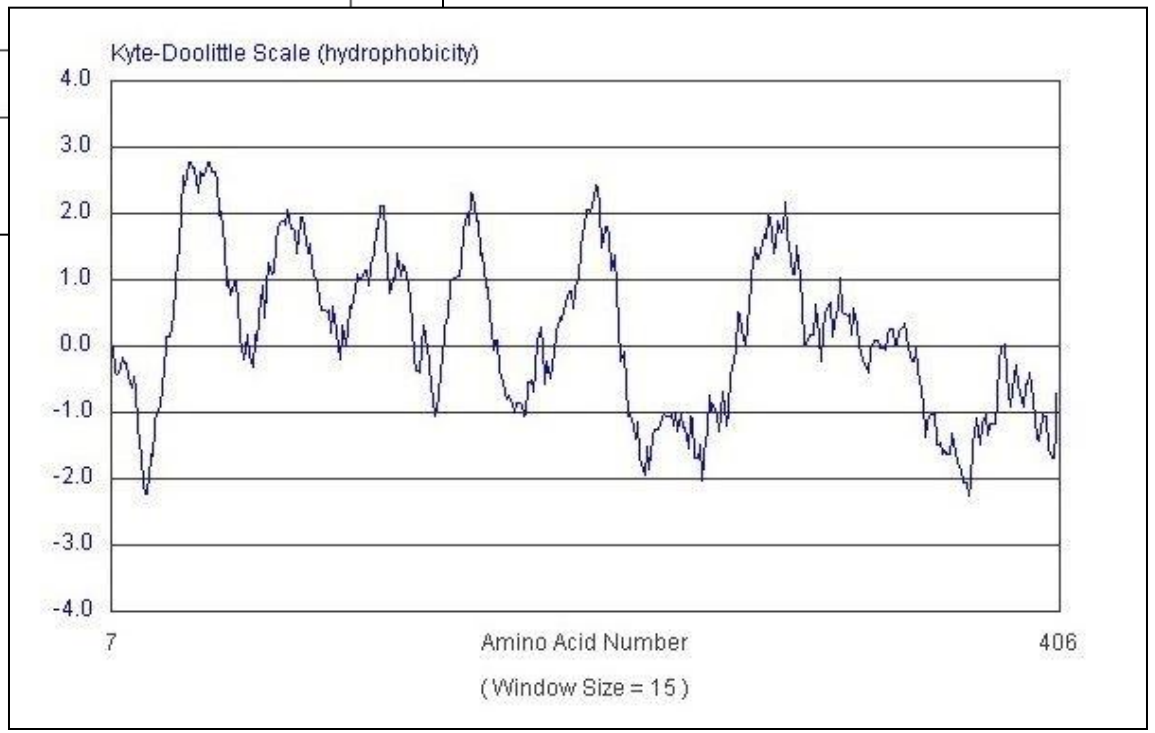
Линейное взвешивание

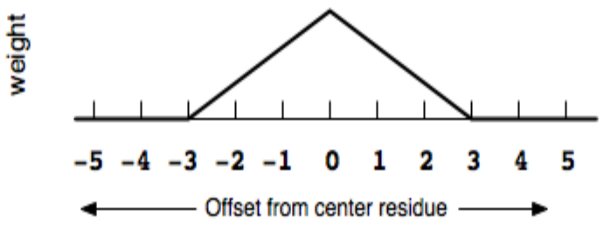


Экспоненциальное взвешивание

Предсказание топологии. Профили гидрофобности







Треугольное взвешивание

Профили гидрофобности

http://www.vivo.colostate.edu/molkit/hydropathy/

«Apologies, but this site was taken down. The programs were written in an older version of Java which is not compatible with most modern browsers and I do not have time to re-code.»

http://gcat.davidson.edu/DGPB/kd/kyte-doolittle.htm

Last Modified: Wednesday, 27 February 2002

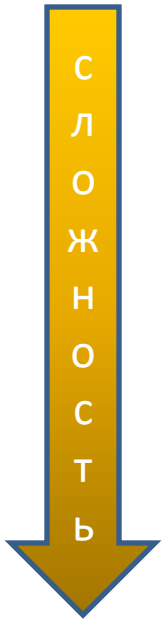
https://web.expasy.org/protscale/

Предсказание структуры белков

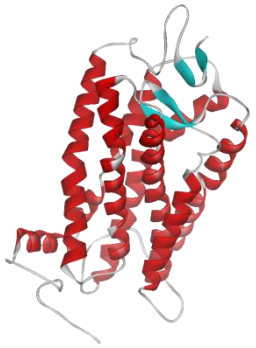
Сворачивание белка в уникальную конформацию наводит на мысль об алгоритме формирования структуры белка по его последовательности, но доказательством полноты и правильности нашего понимания могла бы стать его реализация в виде компьютерной программы...

Методы предсказания структуры по последовательности:

- Предсказание вторичной структуры;
- Моделирование по гомологии;
- Распознавание типов укладки (по известной библиотеке фолдов);
- Априорное предсказание новых типов укладки.



Моделирование на основании гомологии. Алгоритм



Поиск гомологичных белков
с известной структурой (шаблоны)

Выбор подходящего шаблона

Выравнивание последовательности моделируемого белка
с последовательностью шаблона

```
Model:     FVVFVL.FAIC
             :: |: | ||
Template: VIIMVIAFLIC
```

Построение модели

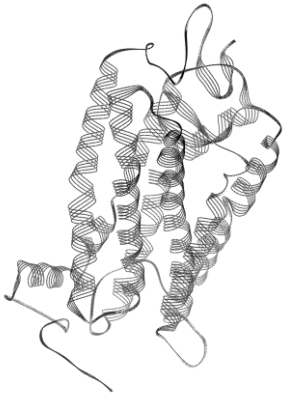
Оценка модели

Нет

Модель подходит?

Да

Ура!

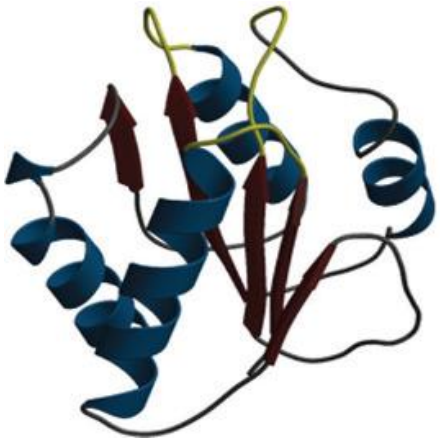


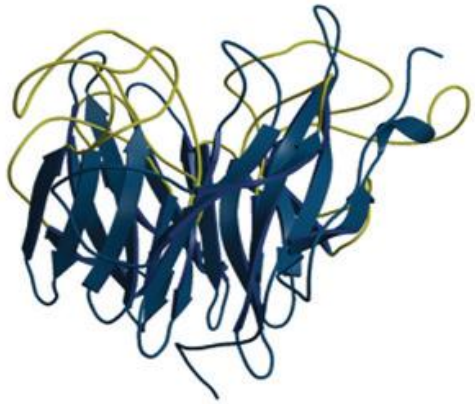
Моделирование на основании гомологии. Методы

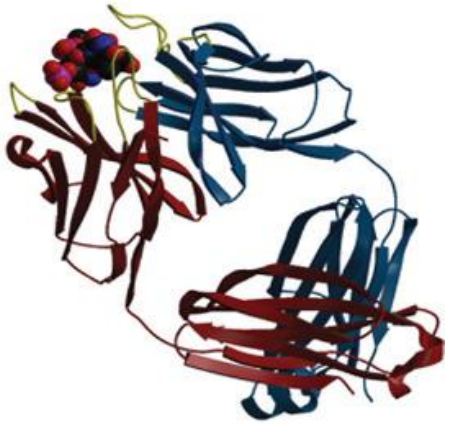
- Сборка модели из «жестких фрагментов»
- Моделирование на основе пространственных ограничений
- Моделирование путем сопоставления сегментов и другие методы

Моделирование петель

- *ab initio*
- путем поиска в базах данных







Сборка модели из «жестких фрагментов»

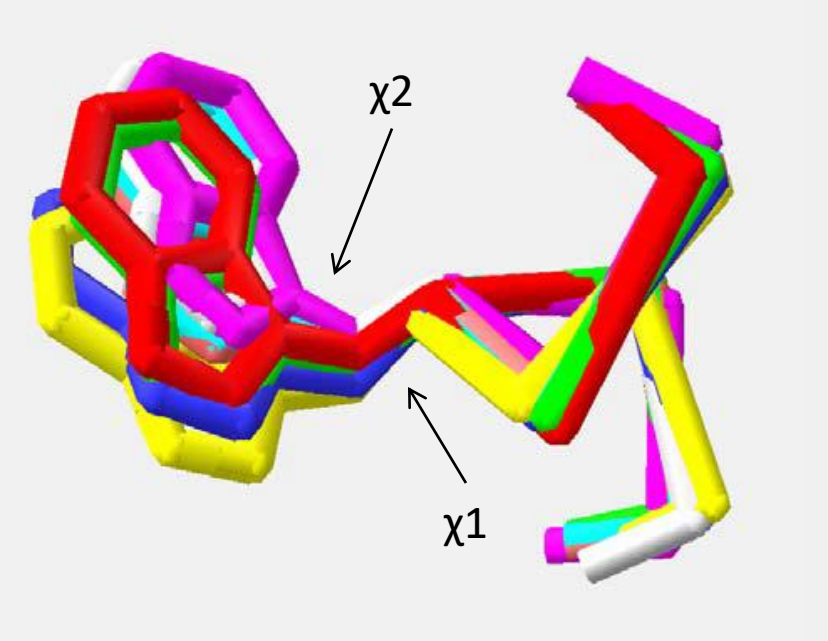
COMPOSER: исторически первый подход к моделированию

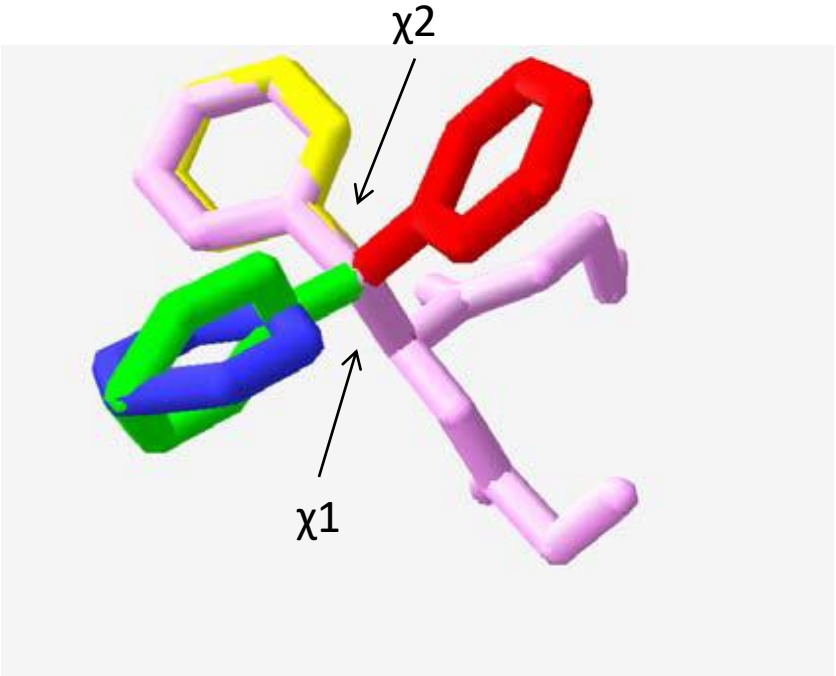
- моделирование в декартовых координатах (Sutcliffe, …, Blundell, 1987)

- Поиск белковых структур с последовательностями, гомологичными моделируемой. Выполнение выравнивание последовательностей, определение положения Cα-атомов консервативных остатков.
- Составление общего шаблона из перекрывающихся структурно консервативных фрагментов (при необходимости).
- Достройка боковых цепей с учетом библиотек ротамеров.
- Достройка петель путем подбора подходящих по геометрии гомологичных фрагментов среди белковых структур.
- Общая оптимизация геометрии.

Библиотеки ротамеров

- Лишь небольшая доля всех возможных конформаций боковых цепей реально наблюдается в экспериментальных структурах
- Конформация боковой цепи зависит от геометрии основной цепи
- Библиотеки ротамеров содержат наборы вероятных конформаций

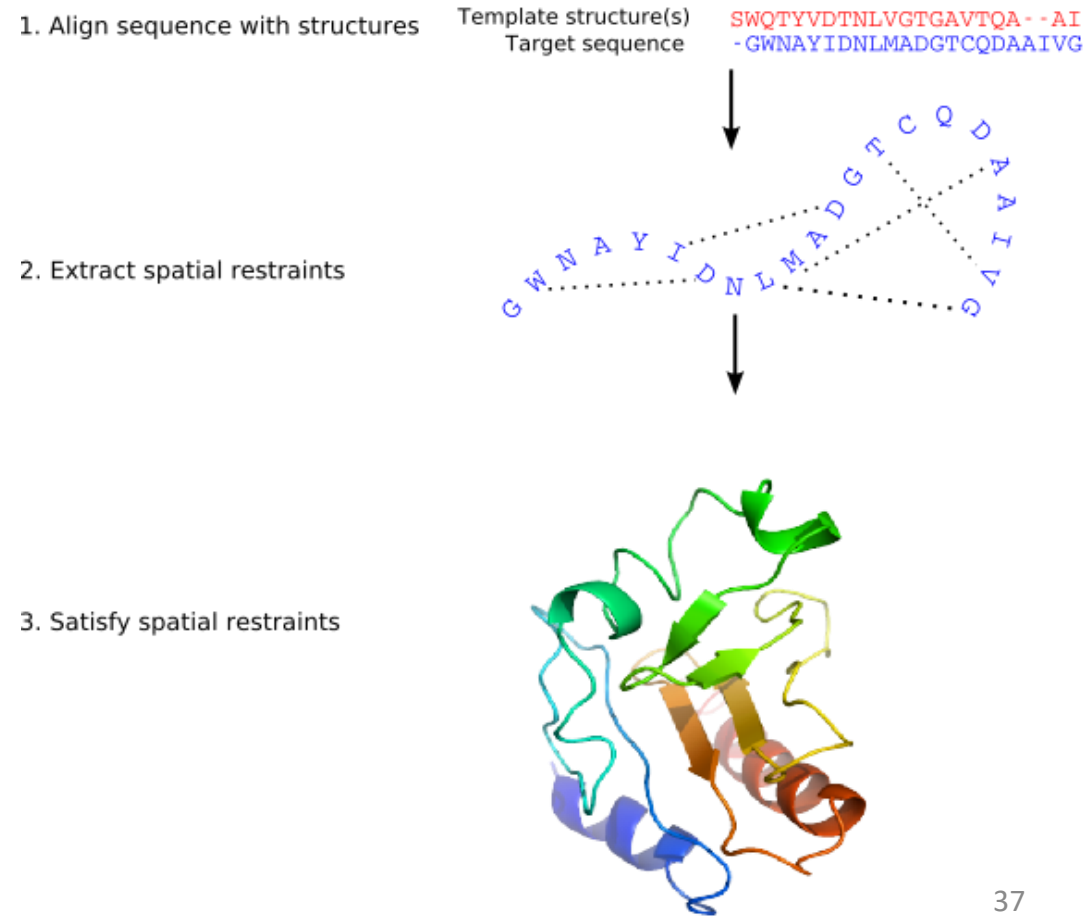




Моделирование на основе пространственных ограничений. MODELLER

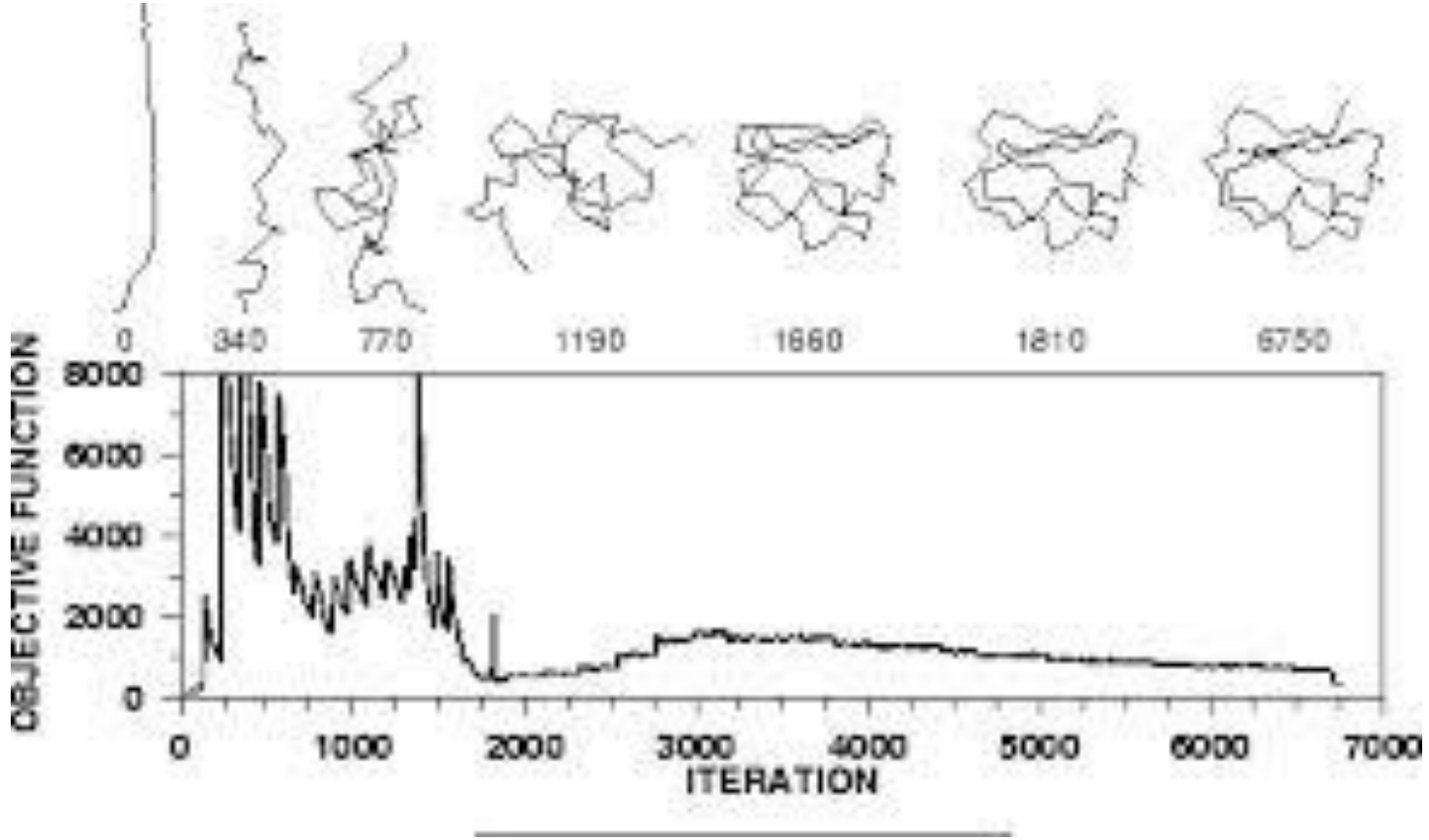
Наиболее распространенный подход к моделированию (Sali & Blundell, 1993) – моделирование во внутренних координатах

- Поиск белковых структур с последовательностями, гомологичными моделируемой.
- Выполнение выравнивания последовательностей.
- Извлечение пространственных ограничений из шаблонов.
- Построение модели с учетом этих ограничений
- Общая оптимизация геометрии.



Моделирование на основе пространственных ограничений. MODELLER

Начиная с распрямленной конформации или конформации шаблона, выполняется учет все более далеких ограничений, чередующийся с минимизацией энергии методом сопряженных градиентов.



Выбор шаблона

Методы, применяемые для сравнения последовательностей:

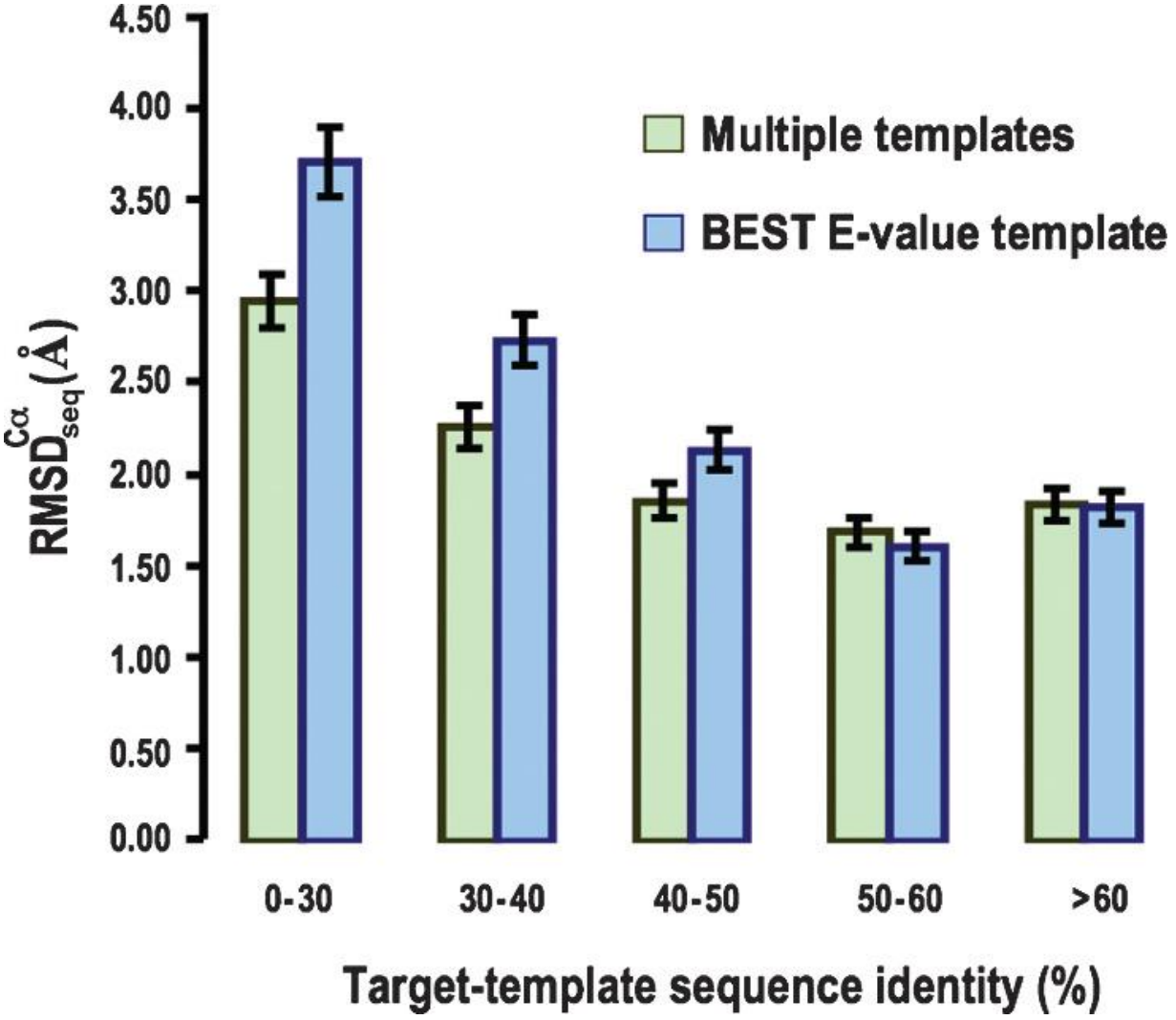
1. Попарное сравнивание моделируемой последовательности с каждой последовательностью из базы данных (FASTA, BLAST).
2. Сравнивание сразу нескольких последовательностей (Clustal, Muscle).
3. Протягивание последовательности через библиотеку пространственных структур.

Факторы, влияющие на выбор шаблона:

1. Высокая идентичность последовательностей.
2. Белки принадлежат к одному подсемейству.
3. Качество экспериментальной структуры (разрешение или количество ограничений на аминокислотный остаток).

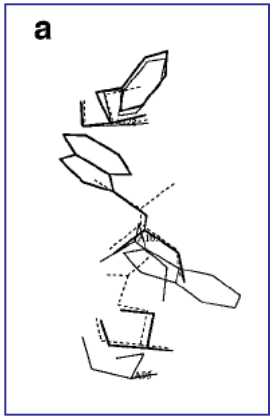
Процент идентичности	Качество выравнивания
> 40%	почти всегда высокое
< 30%	ошибочно выровненные участки
30-40%	«сумеречная зона»

Качество модели: несколько шаблонов

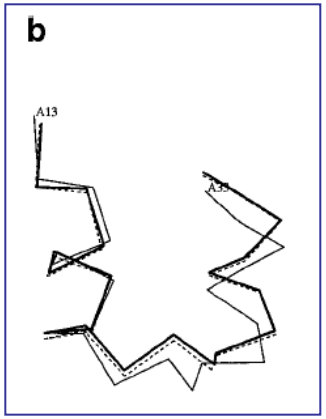


Ошибки построения модели

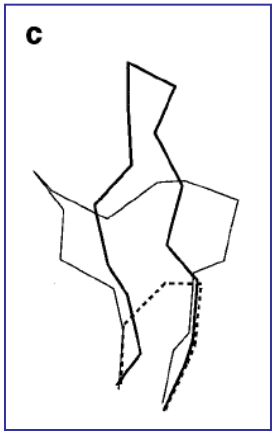
1. Ошибки в ориентации боковых цепей.
Мышиный белок, связывающий ретиноевую кислоту. Тонкая линия – кристалл, толстая линия – модель, пунктир – шаблон (мышиный липид-связывающий белок).



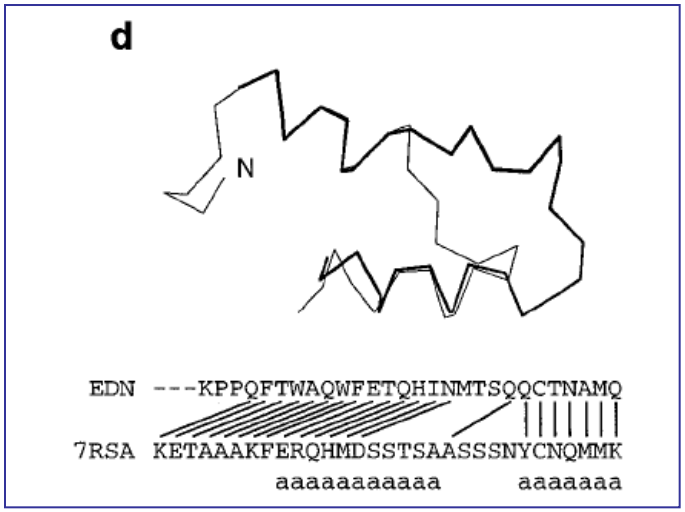
2. Сдвиги в корректно выровненных участках.
Сравнение участка кристаллической структуры мышиного белка, связывающего ретиноевую кислоту, с его моделью и с шаблоном.



3. Ошибки в участках, для которых отсутствует шаблон.
Показан контур Cα атомов остатков 112-117 кристаллографической структуры человеческого эозинофильного нейротоксина (тонкая линия), его модели (толстая линия), и шаблона – рибонуклеаза А (пунктир).



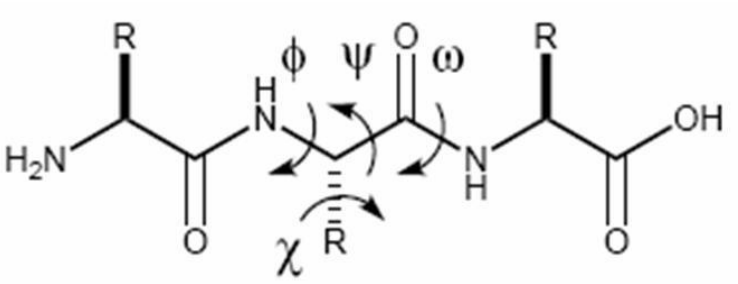
4. Ошибки из-за неправильного выравнивания.
N-концевой участок токсина сравнивается с его моделью. Показан соответствующий участок выравнивания, линии показывают эквивалентные остатки.

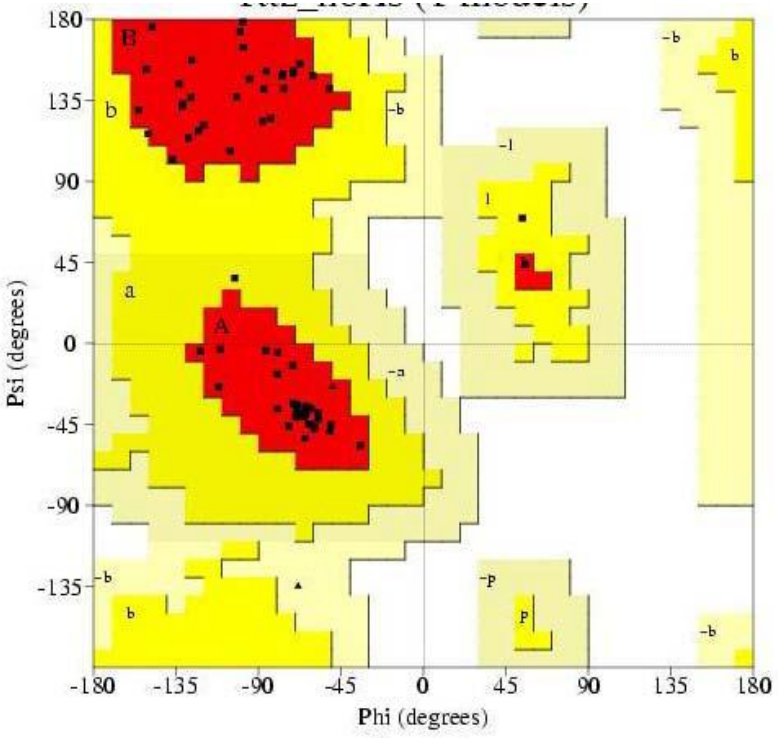


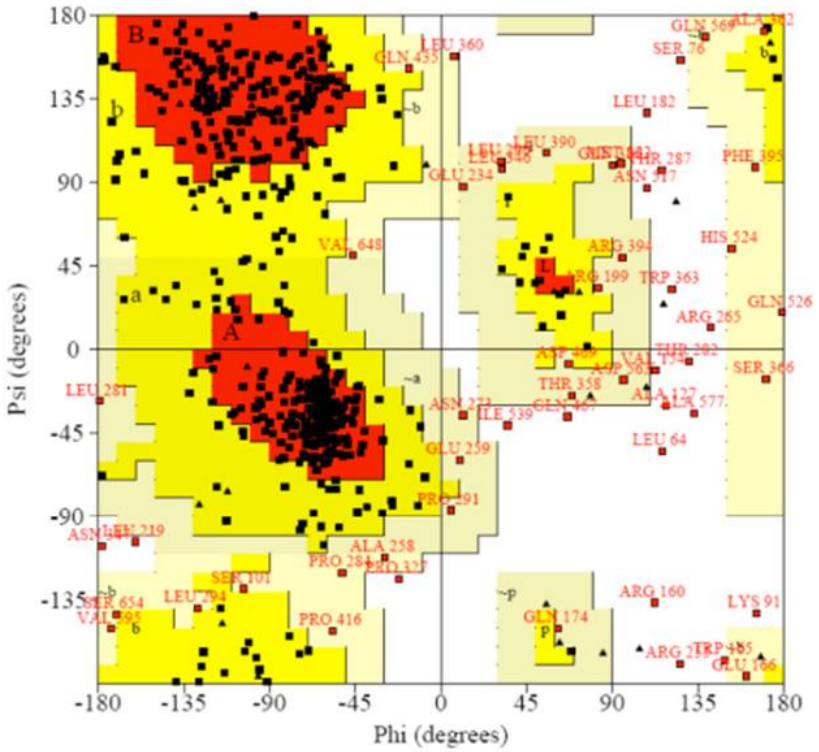
5. Неправильно выбранный шаблон.

Оценка модели. Проверка стереохимии

Карты Рамачандрана



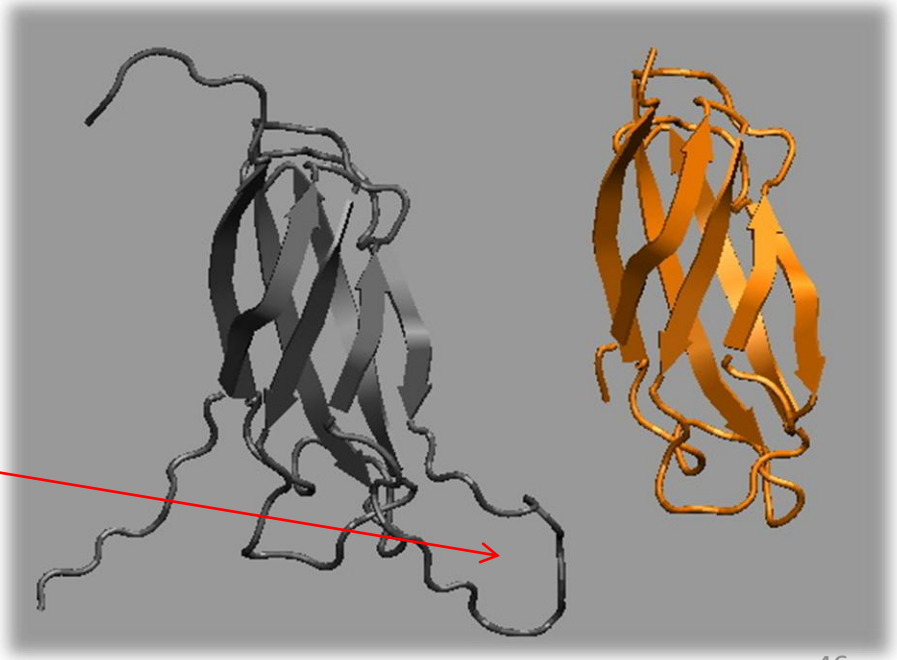




Еще один пример моделирования

```
FN3H    -MQVSDVPTNLEVVAATPTSLLISWYTFTHYG--MNRYYRITYGETGGNS 47
1FNA_A --------RDLEVVAATPTSLLISWDAP-AVT---VRYYRITYGETGGNS 38
               :*************** .          **************

FN3H    PVQEFTVPWINTYTGEPTYADDFKGRFTATISGLKPGVDYTITVYAVTEF 97
1FNA_A PVQEFTVP---------------GSKSTATISGLKPGVDYTITVYAVTGR 73
       ********                ***********************

FN3H    SGTGDFDYPISINYRTLEHHHHHH 121
1FNA_A GDSPASSKPISINYRTEI------ 91
               ********
```

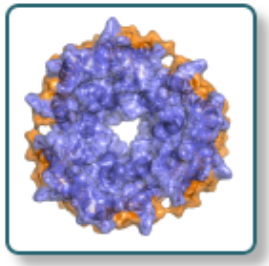
Вставка



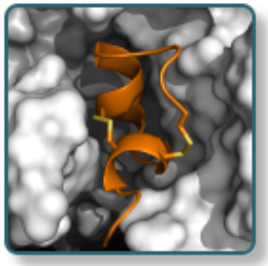
Welcome to ROSIE

Rosetta Online Server that Includes Everyone

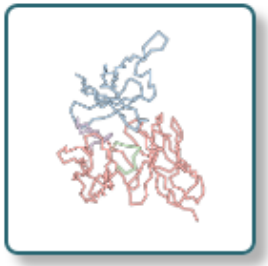
Rosetta Protocols opened for academic users:



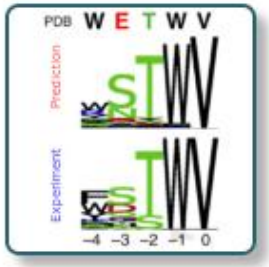
[Mp_lipid_acc]



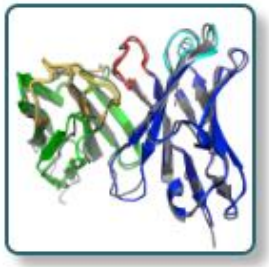
[Tox_dock]



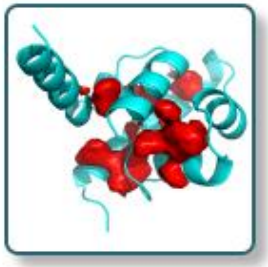
[Snug_dock]



[Sequence_tolerance]



[Antibody]



[Vip]

ROSIE stats (24hrs):

Users:	4,815	+1
Jobs:	31,163	+28
CPU hours:	3,521,032	+5,956

See more info at our About page.

Get Started with ROSIE

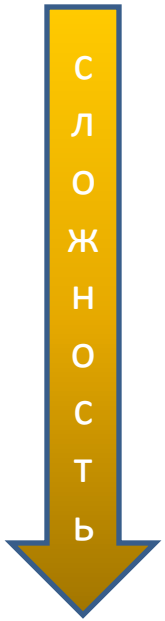
- ROSIE Documentation – Server related documentation and info.
- Rosetta Forums This is a list of forums for Rosetta users to discuss problems with running Rosetta and is monitored by Rosetta developers.

Предсказание структуры белков

Сворачивание белка в уникальную конформацию наводит на мысль об алгоритме формирования структуры белка по его последовательности, но доказательством полноты и правильности нашего понимания могла бы стать его реализация в виде компьютерной программы...

Методы предсказания структуры по последовательности:

- Предсказание вторичной структуры;
- Моделирование по гомологии;
- Распознавание фолда;
- Априорное предсказание новых типов укладки.

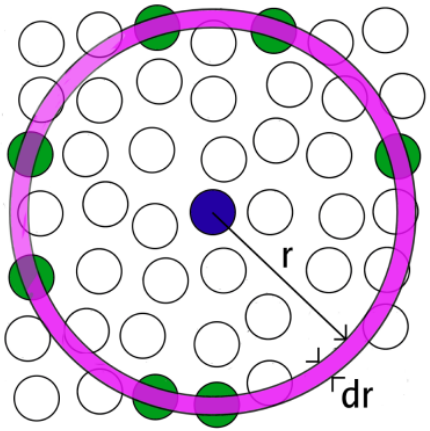


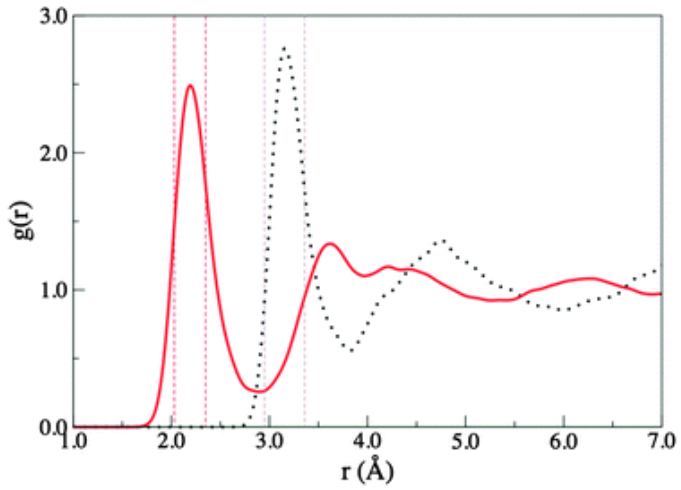
Распознавание фолда. «Протягивание»

Если для последовательности нет гомолога с известной структурой, то, возможно, есть хотя бы структура, подходящая для данной последовательности?
Фактически, нужно примерить данную последовательность на все типы укладки и выбрать наиболее подходящую - **метод «протягивания»** (threading) состоит в построении большого числа грубых моделей для данной последовательности и их последующей экспресс-оценки.

$\Rightarrow$ **нужна функция оценки соответствия последовательности и фолда**.

Например, функции распределения вероятности парных расстояний между остатками (например, по Cβ-атомам) (20x20 штук). Соотнося расстояния в моделях с этими функциями, можно оценить насколько вероятны как эти расстояния, так и модели в целом.



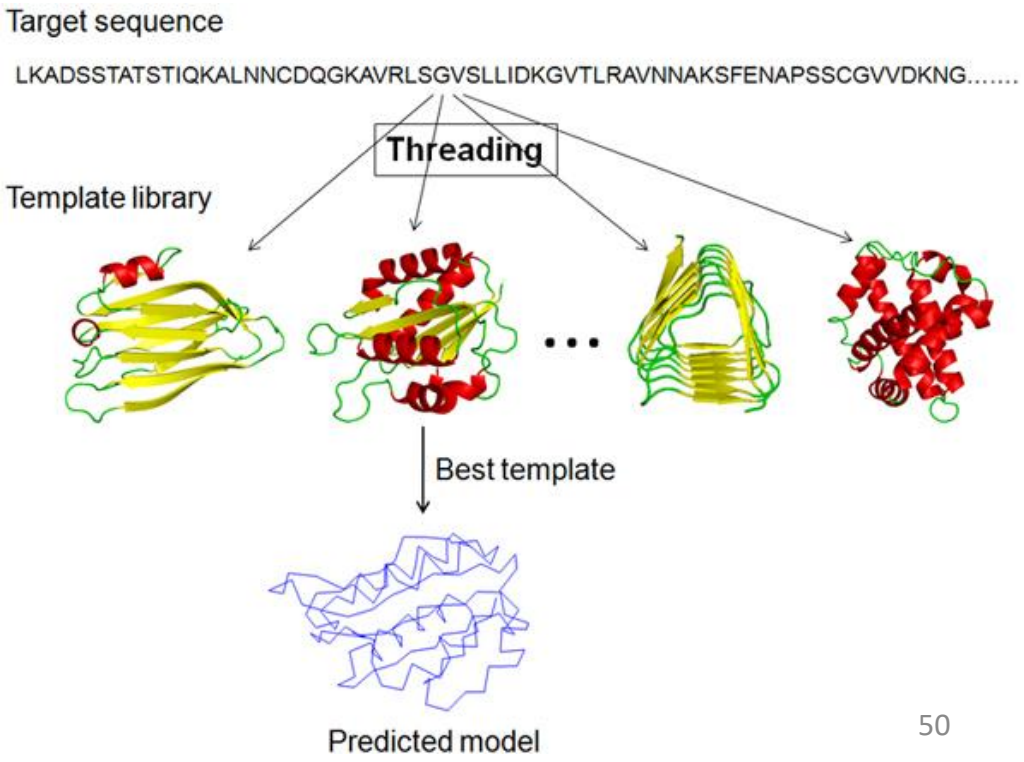


Распознавание фолда. «Протягивание»

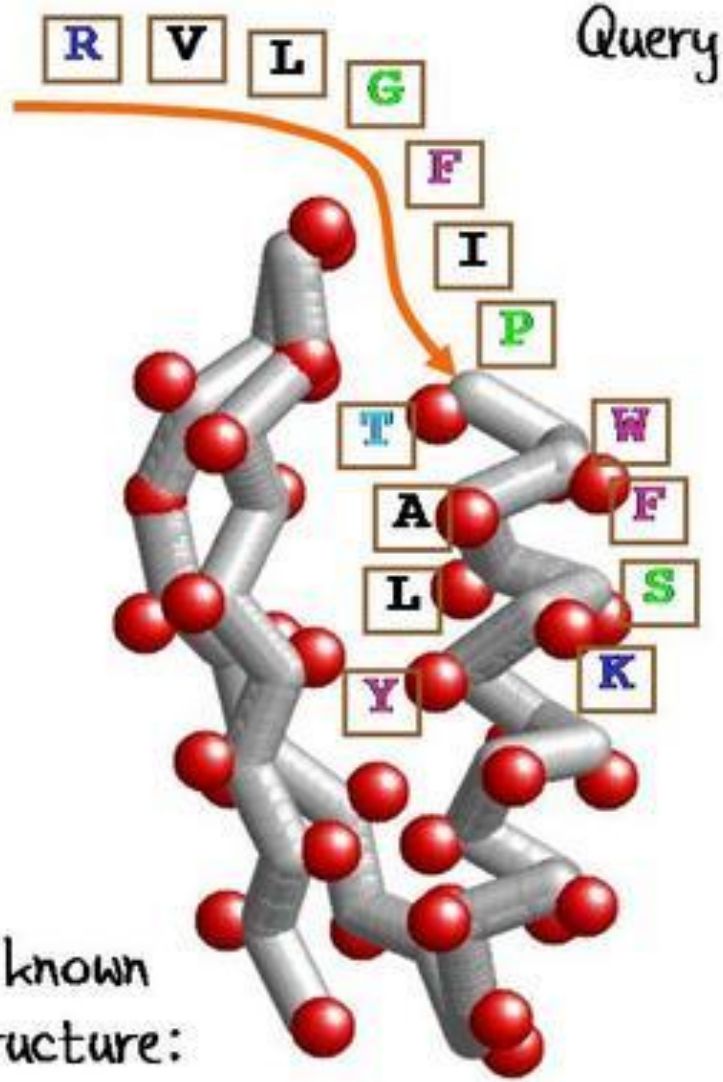
1. Выбор некой известной структуры в качестве потенциального шаблона
2. Генерация всевозможных выравниваний последовательности шаблона с новой последовательностью

```
...IIAWLVKEKKVDVIV...  ...IIAWLVK-EKKVDVIV...  ...IIAWLVKEKKVDVIV...
...NGLELVLDSVLDATF...  ...NGLELVLDSVLDATF...   ...NGLELVLD-SVLDATF...
       **     *               **                        **    *
```

3. Построение и оценка моделей
4. Переход к следующему шаблону (п.1)
5. Сопоставление моделей, построенных по различным шаблонам, и выбор оптимальной



WHAT IS THREADING

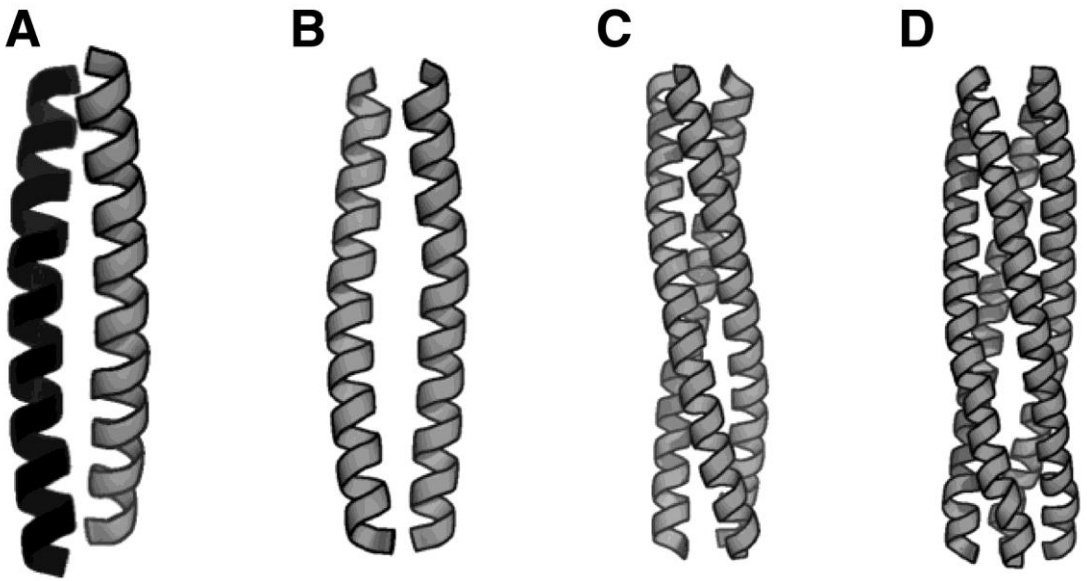


A known structure:

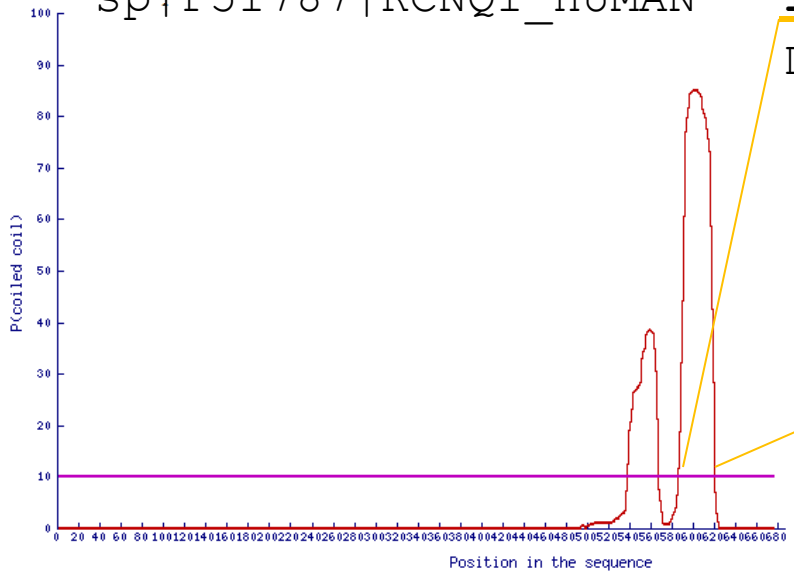
Query Sequence: RVLGFIPTWFALSKY

- Thread the sequence onto the structure.
- Use structural properties to evaluate the fit:
  - Local structure
  - Environment
  - Pairwise interactions

Распознавание фолда. Суперспирали




```
                                590       600       610       620
sp|P51787|KCNQ1_HUMAN           IGARLNRVEDKVTQLDQRLALITDMLHQLLSLH
                                DefgAbcDefgAbcDefgAbcDefgAbcDefGa
```

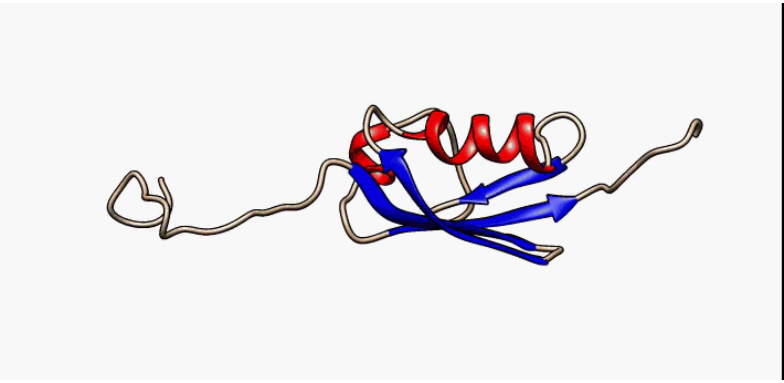


LOGICOIL

Multi-state coiled-coil oligomeric state prediction

Неупорядоченные белки

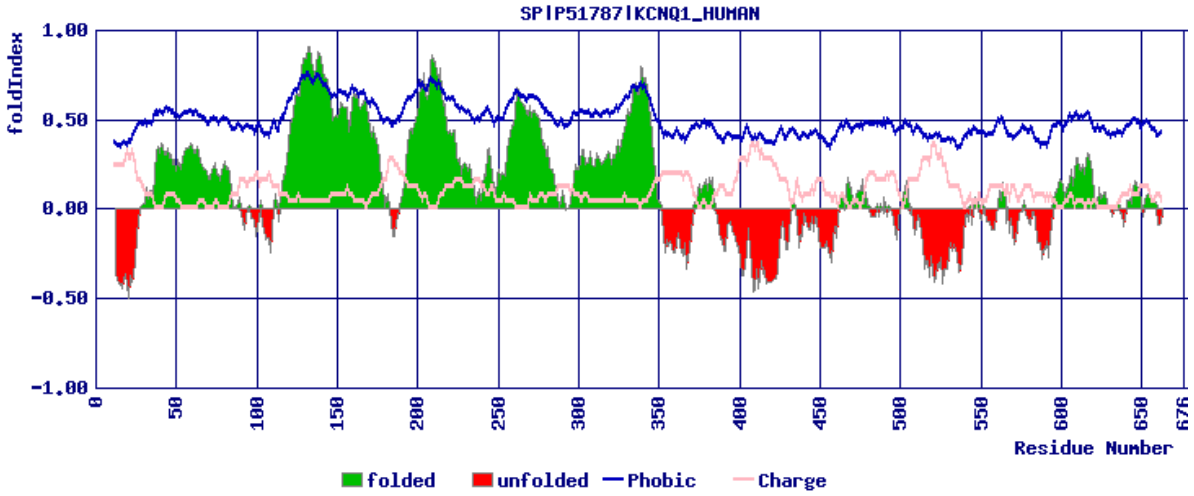
Intrinsically disordered proteins - «нарушители» догмы «структура определяет функцию»

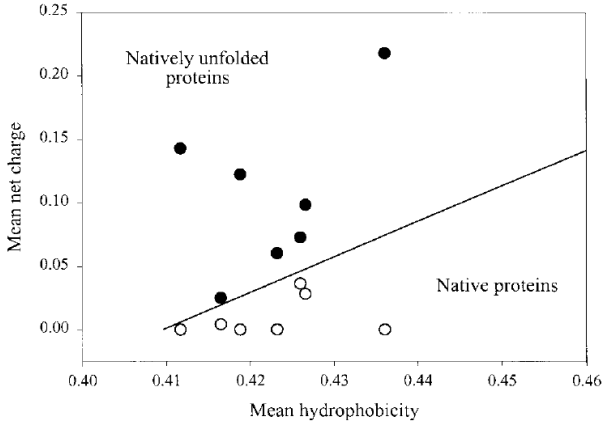


Предсказание неупорядоченных участков по последовательности:
- использование структурных данных (нейронные сети)
- использование свойств аминокислот

$$\langle R \rangle = 2.785 \langle H \rangle - 1.151$$ (Uversky VN, et al. 2000)

FoldIndex http://bip.weizmann.ac.il/fldbin/findex





I-TASSER

Protein Structure & Function Predictions

(The server completed predictions for 390328 proteins submitted by 94188 users from 138 countries)
(The template library was updated on 2018/04/02)